Characterisation of a Gene Trap Integration Marking Hepatic Specification

ALISTAIR JAMES WATT

Thesis presented for the degree of Doctor of Philosophy

University of Edinburgh

1999

To my family

Acknowledgements

I would like to thank my supervisor Lesley Forrester for her advice, endless encouragement and for always making time to discuss the many problems encountered throughout the course of my PhD. Thanks also go to Bill Skarnes, John Ansell and John Mullins for their advice and support as members of my PhD committee.

I would also like to thank past and present members of the Forrester Lab; Gurman Pall, Peter McClive, Phillipe Gabant and Melanie Jackson for the interesting discussions, good laughs and strict adherence to coffee at 11 O'Clock. Thanks to all in the animal house for looking after I114 and Diane Peddie and Tony Coyle for doing all the tails.

Special thanks go to Jonny Boles for Friday nights and Sarah Burl for the frequent coffee breaks during writing up. Finally, I would like to thank Sheena for her love and support.

Abstract

Gene trapping in mouse embryonic stem cells has been used to identify and characterise the function of novel genes. Introduction of a gene trap vector into the genome results in the generation of a fusion between the *lacZ* reporter gene and the endogenous trapped gene. The consequences of this are predicted to be threefold: (i) expression of the reporter gene will be controlled by the promoter and enhancer elements of the endogenous gene and will therefore mirror endogenous gene expression; (ii) the generation of a fusion transcript allows endogenous gene sequence to be cloned by 5'RACE-PCR and (iii) the insertion of the gene trap vector can disrupt the function of the endogenous gene.

This work describes the characterisation of one specific gene trap integration, I114 that has not behaved entirely as predicted but has none-the-less identified an early marker of hepatic specification. The reporter activity profile associated with the I114 gene trap integration is restricted to the definintive endoderm marking the ontogeny of the foetal liver. Reporter activity is observed as early as the 9 somite stage (8.0-8.5dpc) in endodermal cells of the foregut in the region destined to form the liver diverticulum and is restricted to the hepatic lineage until late gestation. Comparison of I114 reporter activity with AFP identifies I114 reporter activity as being the earliest, most specific marker of liver organogenesis identified to date.

Breeding of the gene trap integration to homozygosity reveals no overt phenotype but its unique pattern of expression prompted us to clone the endogenous sequence. This has proven to be more complex than predicted as integration of the gene trap vector results in the production of two fusion transcripts. The most abundant (Group I) fusion transcript is ubiquitously expressed. No reporter activity is produced from this fusion transcript as gene trap vector splicing to this sequence places translation of the *lacZ* gene out-of-frame. The second, less abundant (Group II) fusion transcript is expressed exclusively in the liver during embryogenesis and is predicted to produce reporter activity. The cloning and sequencing of both the genomic sequence and the endogenous gene (*gtar* - **g**ene **t**rap **a**nkyrin **r**epeat) associated with the Group I and II fusion sequences has revealed that they represent different exons of *gtar*. Expression of the Group I and Group II sequences independent of the gene trap vector mimics that of the I114 fusion transcripts. Furthermore, expression of these sequences in I114 homozygous tissues indicates that the insertion of the gene trap vector has failed to disrupt the expression of *gtar*. Expression of the liver specific exon of *gtar* is postulated to be a consequence of either a separate promoter immediately upstream of the exon or alternative splicing.

Contents

Declaration
Dedication
Acknowledgements
Abstract
Contents

Chapter 1: INTRODUCTION
1.1. Mutational Analysis of the Mouse 1
1.1.1. Random Mutagenesis 1
1.1.2. Directed Mutagenesis 2
1.1.3. Insertional Mutagenesis 4
1.2. Entrapment Technology 5
1.2.1. Entrapment In A Variety Of Model Systems 5
1.2.2. Entrapment In The Mouse 7
1.3. Entrapment in ES Cells 8
1.3.1. Enhancer Trappping 8
1.3.2. Promoter Trapping 9
1.3.2.1. Entrapment Aids Gene Identification 10
1.3.2.2. Reporter Activity Reflects Endogenous Gene Expression 11
1.3.2.3. Vector Insertion Is Mutagenic 12
1.3.2.4. Promoter Trapping Technology 13
i/ Vector design 13
ii/ Reporter gene 14
iii/ Vector delivery 16
iv/ Entrapment is a random event 19
v/ PolyA trapping 20
vi/ Gene trapping and site specific recombination 21
1.3.2.5. Prescreening Gene Trap Events 22
i/ Sequence 22
ii/ Secretory trap vectors 23
iii/ Reporter expression 24
iv/ *In vitro* prescreening 25
1.3.2.6. Summary 26

1.4. Liver Development 28
1.4.1. Overview Of Liver Development 28
1.4.2. Tissue Interactions Mediating Liver Development 29
1.4.3. Hepatic Markers 31
1.4.4. Molecular Basis of Hepatic Determination 32
1.4.5. Mutational Analysis of Liver Development 34
1.4.6. Summary 36
1.5. Experimental Approach 37
Chapter 2: MATERIALS AND METHODS 38
2.1. Molecular Biology Methods 38
2.1.1. General Cloning Techniques 38
2.1.1.1. Gel Purification of DNA 39
2.1.1.2. Ligations 40
2.1.1.3. Transformation of Bacterial Cells 40
2.1.1.4. Screening Transformants 41
2.1.2. Isolation of Nucleic Acids 42
2.1.2.1. Plasmid Preparation 42
2.1.2.2. PAC Preparation 43
2.1.2.3. Isolation of Genomic DNA 44
2.1.2.4. Isolation of High Molecular Weight Genomic DNA 45
2.1.2.5. Isolation of RNA 45
2.1.3. Analysis of High Molecular Weight DNA 46
2.1.3.1. Digestion of Genomic DNA Agarose Plugs 46
2.1.3.2. Pulse Field Gel Electrophoresis 47
2.1.4. Nucleic Acid Transfer to Membranes 48
2.1.4.1. Southern Blotting 48
2.1.4.2. Dry Blotting 48
2.1.4.3. Alkali Blotting 49
2.1.4.4. Dot Blotting 49
2.1.4.5. Northern Blotting 50
2.1.5. Radiolabelling Probes 50
2.1.5.1. Random Priming Probes 50
2.1.5.2. End-labelling Oligonucleotide Probes 51
2.1.5.3. Table of Probes 51
2.1.6. Hybridisation Conditions 52
2.1.7. DNA Sequencing 53

2.1.7.1. Manual Sequencing 53
2.1.7.2. Automated Cycle Sequencing 53
2.1.7.3. Direct Sequencing of RACE-PCR Products 54
2.1.7.4. Sequencing Primers 56
2.1.8. Polymerase Chain Reaction (PCR) 56
2.1.8.1. Rapid Amplification of cDNA Ends-PCR (5'RACE-PCR) 56
2.1.8.2. Reverse Transcriptase-PCR (RT-PCR) 62
2.1.9. RNase Protection Assay 63
2.1.10. cDNA Library Screening 65
2.2. Protein Analysis 69
2.2.1. Protein Preparation 69
2.2.2. SDS-Polyacryamide Gel Electrophoresis (SDS-PAGE) 69
2.2.3. Coomassie Staining 70
2.2.4. Immunoblotting 70
2.2.5. Immunodetection 70
2.3. ES Cell Culture 72
2.3.1. Reagents 72
2.3.2. Thawing ES Cells 72
2.3.3. Passsage and Expansion of ES Cells 73
2.3.4. Freezing ES Cells 73
2.3.5. Selection of Retinoic Acid-Responsive Gene Trap Cell Lines 74
2.4. Histology 75
2.4.1. Maintenance of Animals 75
2.4.2. Preparation of Specimens for Histology 75
2.4.3. Cryostat Sectioning 75
2.4.4. Staining Embryos and Cryostat Sections for β-gal Activity 76
2.4.5. Haematoxylin and Eosin Counterstaining 76
2.4.6. Microscopy and Photography of Specimens 77

Chapter 3: RESULTS 78
Characterisation of I114 Reporter Expression Profile and Phenotype
3.1. Introduction 78
3.2. I114 Gene Trap Cell Line 78
3.3. I114 Expression Analysis 79
3.3.1. Embryonic Expression 79
3.3.2. Adult Expression 81

3.4. Comparison of I114 Reporter Activity With AFP Expression 82
3.4.1. Embryonic Expression 82
3.4.2. Adult Expression 83
3.5. I114 Reporter Activity During Tumourigenesis 84
3.6. I114 Breeding Analysis 85
3.7. Discussion 87

Chapter 4: RESULTS 93
Molecular Characterisation of the I114 Gene Trap Integration
4.1. Introduction 93
4.2. I114 Molecular Analysis 93
4.2.1. Identification of I114 Fusion Transcripts Using 5'RACE-PCR 94
4.2.2. Analysis of RACE Clone Sequence 96
4.3. Expression Analysis of Group I and II Fusion Transcripts 97
4.4. LacZ Protein Expression Analysis 102
4.5. Discussion 103

Chapter 5: RESULTS 108
Genomic Characterisation of the I114 Gene Trap Integration
5.1. Introduction 108
5.2. Chromosomal Location of the I114 Gene Trap Integration 108
5.3. Genomic Structure Analysis of the I114 Vector Copies 109
5.4. Cloning and Analysis of Group I and II Genomic DNA 110
5.5. Summary 114

Chapter 6: RESULTS 115
Isolation and Characterisation of Gene Trap Ankyrin Repeat (*gtar*)
6.1. cDNA Library Screening Using Group I Sequence 115
6.2. Analysis of the Group I Sequence 116
6.3. Protein Sequence Analysis of *gtar* 117
6.4. Isolation of *gtar* Splice Variants 119
6.5. Genomic structure of *gtar* 120
6.6. Expression Analysis of *gtar* 121
6.7. Cloning and Sequencing of the Endogenous Group II Transcript 124
6.8. Discussion 127

Chapter 7: CONCLUDING REMARKS 138

Appendix I 141
Appendix II 142
Appendix III 143
Appendix IV 144
Abbreviations 147
List of Figures 149
List of Tables 150

References 151

Chapter 1

INTRODUCTION

1.1. MUTATIONAL ANALYSIS OF THE MOUSE

Mutational analysis has revolutionised our understanding of the processes involved in embryonic development. Phenotype driven screens in invertebrate systems such as *Drosophila melanogaster* (Nüsslein-Volhard, 1984) and *Caenorhabditis elegans* (Kempheus, 1988) have been successful in identifying many of the molecules and pathways essential for correct embryonic patterning in these organisms. Comparable saturation scale mutagenesis has been performed in the Zebrafish *Danio rerio*, providing an excellent resource for the analysis of vertebrate embryogenesis (Driever *et al.*, 1996; Haffter *et al.*, 1996). However, the identification of the genes associated with the phenotypes is hindered by the relatively poorly characterised Zebrafish genome. The mouse provides the ideal mammalian system for mutational analysis with well characterised inbred strains, over 1000 existing mutant loci and an increasingly well defined genome. Furthermore, transgenic technology allows for a vast range of modifications to the mouse genome.

1.1.1. Random Mutagenesis

The mouse has an existing pool of over 1000 spontaneous, chemical or irradiation induced mutants (Doolittle *et al.*, 1996). Moreover, mutagenic screens to identify additional mutants for study have been undertaken in mice using X-rays or chemical mutagens such as ENU and chlorambucil (Rinchik, 1991). Initially these screens have focused on the identification of novel mutants within small, phenotypically defined genomic regions (Shedlovsky *et al.*, 1988; Rinchik, 1991) or to the isolation of

mutations with specific phenotypes (Shedlovsky *et al.*, 1993; Hotz Vitaterna *et al.*, 1994). Currently the MRC Mammalian Genetics Unit at Harwell is undertaking a large scale ENU mutagenesis programme in the mouse (http://www.mgu.har.mrc.ac.uk/mutabase/). Two approaches are being used; a genome wide screen for dominant mutations and a targeted screen for recessive mutations within a deletion region on chromosome 13. Essential to both of these approaches is that phenotype screening of each mutant will be performed using a defined protocol of tests which will identify abnormalities ranging from embryonic lethality through to subtle behavioural abnormalities (Rogers *et al.*, 1997). Increasingly refined mapping and gene identification techniques are aiding the significant task of isolating the mutated gene responsible for individual phenotypes. Interspecific backcross panels and the use of simple sequence length polymorphisms (SSLP) are defining a high resolution linkage map of the mouse genome allowing for more accurate mapping of mutant loci (Dietrich *et al.*, 1996; Avner *et al.*, 1988; Bedell *et al.*, 1997). Positional cloning and the candidate gene approach have subsequently been successful strategies for the identification of the mutated gene. The genes responsible for many developmentally significant mutants have been isolated using these techniques, including *T*, *inv*, and *sm* by positional cloning (Hermann *et al.*,1990; Morgan *et al.*, 1998; Sidow *et al.*, 1997) and *W*, *Steel*, and *un* by identifying candidate genes (Reith *et al.*, 1990; Huang *et al.*, 1990; Balling *et al.*, 1988).

A significant limitation of the random mutagenesis approach in the mouse compared to non-mammalian systems is the considerable resources needed to maintain the breeding numbers necessary to define and map mutations.

1.1.2. Directed Mutagenesis

The development of transgenic technology, in particular targeted homologous recombination in ES cells (Capecchi, 1989) provides a complementary approach to

phenotype driven screens with studies orientated towards gene identification and targeted mutational analysis.

Diverse approaches have been employed to identify novel genes that play important roles in murine development. For example, the identification of homologues of pattern formation genes discovered in genetically more accessible organisms such as *Drosophila* has been hugely influential in isolating gene families essential in murine embryogenesis (Kessel and Gruss, 1990). However, such approaches will not identify genes that are mammalian specific.

Many genes involved in mammalian signalling cascades initially identified, for example, by their oncogenic potential or effects *in vitro*, have shown restricted developmental expression patterns and subsequently displayed developmental phenotypes when disrupted *in vivo* (McMahon, 1992; Sanford *et al.*, 1997; Yamaguchi *et al*, 1994; Schmidt *et al.*, 1995).

Strategies for the *de novo* isolation of developmental genes based on expression profile include subtractive hybridisation (Lee *et al.*,1991; Harrison *et al.*, 1995), differential display (Liang and Pardee, 1992; Welsh *et al.*,1992) and the identification of factors binding upstream of tissue specific genes (Costa *et al.*, 1989). These technologies allow for more focused searches, enriching for genes with desired expression patterns and has been successful at identifying developmentally important genes (Dunwoodie *et al.*, 1998; Chen *et al.*, 1994a; Lai *et al.*, 1993).

Expressed sequence tag (EST) databases of randomly sequenced cDNA clones from a wide spectrum of developmental stages and tissues provides another potentially important *de novo* source of developmental genes (Adams *et al.*, 1995; Boguski *et al.*, 1993).

ES cell mediated gene knock-out studies have allowed for the function of many of the genes isolated by these techniques to be examined *in vivo*. A catalogue of 327 individual gene knock-outs was first published by Brandon *et al.*, (1995). From the beginning of 1999 this has grown into a database of over 1400 single and double knockouts as well as a number of insertional and classical mutants (http://www.

biomednet.com/db/mkmd). The targeted disruption of individual genes has revolutionised our ability to assess gene function during embryogenesis. However, information regarding genomic gene structure is needed to construct targeting vectors and there is no guarantee that individual gene knock-outs will produce a developmentally relevant phenotype if gene function is redundant among family members.

1.1.3. Insertional Mutagenesis

Another source of informative developmental mutants has arisen from the large numbers of transgenic mouse strains produced over the last 15 years. Around 5% of transgenic and retroviral insertions are believed to be mutagenic. These phenotypes are not attributable to the transgene itself, but due to the insertion of the exogenous DNA interfering with normal gene function either directly or by inducing chromosomal rearrangements and deletions (Rijkers *et al.*, 1994; Gridley *et al.*, 1987). The potential benefit in studying phenotypes arising from insertional mutants is that the exogenous DNA provides a molecular handle on the mutated locus eliminating extensive breeding strategies and potentiating identification of the effected gene. An excellent example of an insertional mutant leading to the isolation of a developmentally important molecule is from the retrovirally induced 413-d transgenic line. Mice homozygous for the retroviral insertion fail to gastrulate correctly showing impairment of embryonic mesoderm formation (Iannaconne *et al.*, 1992; Conlon *et al.*, 1991). The phenotype was subsequently shown to be associated with the disruption of *nodal* a member of the transforming growth factor -β (TGF-β) superfamily of secreted molecules (Zhou *et al.*, 1993). However, identification of the disrupted transcript in insertional mutants has proven a laborious process borne out by how few genes associated with insertional mutants have been identified to date (Rijkers *et al.*, 1994). The high copy number of transgene insertions and transgene induced rearrangements and deletions of endogenous

DNA at the insertion site all complicate the molecular analysis of the mutant locus.(Gridley *et al.*, 1987; Jaenisch, 1988).

1.2. ENTRAPMENT TECHNOLOGY

This section will focus on gene entrapment technology as a means of inducing and identifying developmental mutants. The overall principles of gene entrapment and its use in various experimental systems will serve as an introduction to entrapment in murine ES cells.

Gene entrapment uses insertional mutagenesis in a systematic and reproducible manner to identify genes and their transcriptional control elements. Entrapment vectors consist of a reporter gene which is introduced into the genome where its expression is dependent on insertion either adjacent to *cis*-acting transcriptional control elements (enhancer traps) or into a gene (promoter traps) (Figure 1.1; Skarnes, 1990). Enhancer trap vectors consist of a reporter gene driven by a minimal promoter. Expression of the reporter gene requires the activity of the endogenous enhancer elements (Figure 1.1A; Bellen *et al.*, 1990; Allen *et al.*, 1988). In promoter trap constructs the reporter gene lacks a promoter with its expression dependent on vector insertion into a transcriptional unit for endogenous promoter activity (Figure 1.1B; Brenner *et al.*, 1989; Gossler *et al.*, 1989). Entrapment vectors therefore, although potentially mutagenic, can identify genes based solely on reporter expression. This allows non-essential genes to be isolated, a class of gene not accessible to phenotype driven screens.

1.2.1. Entrapment In A Variety Of Model Systems

The principles of entrapment technology have been applied to many different model systems. The first entrapment vectors were used in *Escherichia coli* and were based on the bacteriophage *Mu* which contained ampicillin and *lacZ* as reporter genes. Random integration of the promoter trap vectors into the bacterial genome tagged

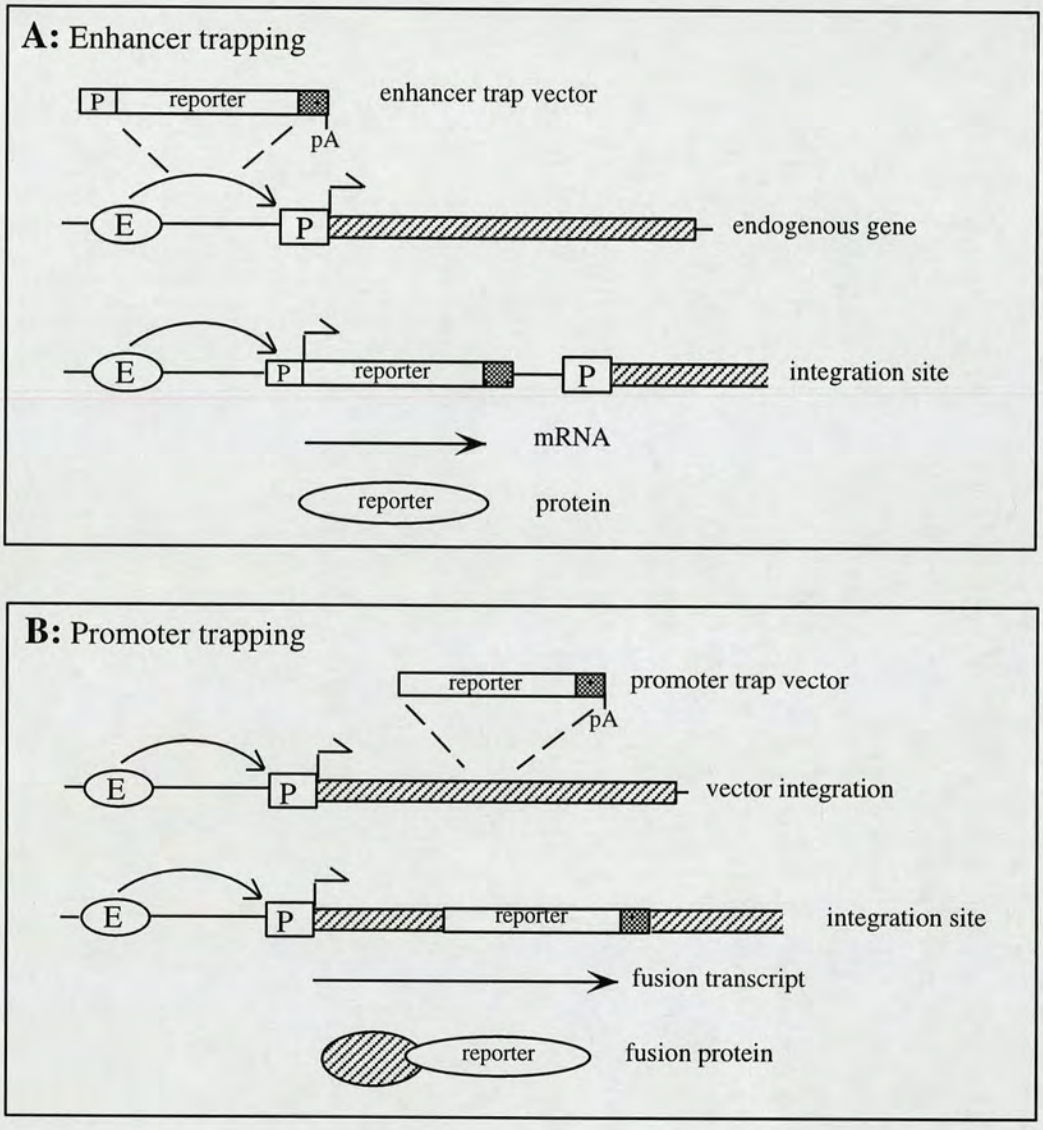


Figure 1.1 Mode of action of the two basic entrapment vector systems

A: Enhancer trapping

Enhancer trap vectors consist of a reporter gene under the control of a minimal promoter. Insertion adjacent to endogenous enhancer elements drives expression of the reporter gene.

B: Promoter trapping

Promoter trap vectors essentially consist of a reporter gene. Insertion into endogenous transcription units is necessary for reporter activity. Reporter expression is driven by the endogenous promoter and enhancer elements and therefore mimics endogenous gene expression.

E, enhancer elements; P, promoter element; pA, polyadenylation signal.

transcriptional promoters for functional and mutational studies (Casabadan *et al.*, 1979; Bellafato *et al.*, 1984). Similar screens have been carried out using mouse embyronic fibroblasts and embryonic carcinoma (EC) cells using both enhancer and promoter traps to identify transcriptionally active loci in culture (Bhat *et al.*, 1988; Brenner *et al.*, 1989; Von Melchner *et al.*, 1990). Functional studies on cell lines from these screens have isolated genomic regions modulating reporter expression in response to differentiation (Bhat *et al.*, 1988) or cell cycle arrest (Brenner *et al.*, 1989).

The application of entrapment technology to developmental biology has been most successful in the fruit fly *Drosophila melanogaster*, with limited studies having been undertaken in the nematode *Caenorhabditis elegans* (Hope, 1991) and in slime moulds (Fey and Cox, 1997). In *Drosophila*, large scale screens have been carried out to identify and mutate genetic elements conferring temporally or spatially restricted expression patterns on the reporter gene during embryogenesis (Bellen *et al.*, 1990). The rationale that tissue restricted gene expression is a prerequisite of embryonic patterning makes the reporter expression conferred by endogenous elements an important factor in assessing potential gene function. Enhancer traps are based on P-element transposons which lack transposase activity but contain the *lacZ* reporter gene that encodes for β-galactosidase (β-gal), controlled by a ubiquitous minimal promoter. β-gal activity provides an excellent reporter for these studies as it can be easily monitored in the embryo using the chromogenic substrate X-gal (O'Kane and Gehring, 1987). The transposons randomly integrate into the fly genome when transposase is provided in *trans* . Heritable, potentially mutagenic insertions are generated where β-gal activity in the embryo is regulated by adjacent regulatory elements (O'Kane and Gehring, 1987). In parallel large scale studies, 60-65% of insertional lines generated showed restricted reporter expression during development in an extensive range of tissues with 10-20% of transposon insertions being lethal when bred to homozygosity (Bier *et al.*,1989; Bellen *et al.*, 1989). In a sub-set of these lines examined, the reporter expression correlated exactly with that of the endogenous transcripts found adjacent to the insertion sites (Wilson *et al.*, 1989). Enhancer trapping therefore provides a

powerful tool for identifying and mutating genes involved in *Drosophila* development. In addition, these lines provide a variety of cell-type and position dependent markers with which to further study *Drosophila* embryogenesis. Thousands of mutagenic P-element insertions are maintained within the Berkeley *Drosophila* Genome Project (BDGP) as a means of linking the large amount of gene sequence data being produced to gene function (Spradling *et al.*, 1995).

1.2.2. Entrapment In The Mouse

The production of transgenic animals via the pronuclear injection of exogenous DNA into fetilised mouse eggs was used to assess the feasibility of using entrapment technology *in vivo*. Injection of an enhancer trap vector with *lacZ* driven by the Herpes Simplex Virus Thymidine Kinase promoter (HSV-tk) produced 52 embryos with an integrated transgene of which 11 gave a range of distinct reporter expression patterns. In addition, from 20 transgenic lines containing this construct, 5 showed unique, heritable reporter expression profiles (Allen *et al.*, 1988). However, the laborious nature of producing transgenic mice with enhancer trap events (only 11 embryos from 200 pronuclear injections showed reporter expression (Allen *et al.*, 1988)) coupled with the problems associated with identifying adjacent DNA from transgenic insertion sites (Gridley *et al.*, 1987) makes it impractical to use this as a routine entrapment system in the mouse.

1.3. ENTRAPMENT IN ES CELLS

The viability of routinely performing developmental entrapment screens in the mouse improved with the development of embryonic stem (ES) cell technology. ES cells are cultured from the inner cell mass of blastocysts stage embryos (Martin, 1981; Evans and Kaufman, 1981). In culture, ES cells are amenable to a range of genetic manipulations and can be reintroduced into intact blastocysts where they can contribute to all cell lineages in chimaeric embryos including the germline (Bradley *et al.*, 1984; Robertson *et al.*, 1986; Thomson *et al.*, 1989). Therefore, ES cells allow for the genetic alteration of the mouse genome in culture and the consequences can be examined *in vivo*. Entrapment constructs can be introduced into the mouse genome using ES cells and rare integration events into transcriptionally active loci can be screened *in vitro* (Gossler *et al.*, 1989; Friedrich and Soriano, 1991). ES cell lines with selected entrapment events can be produced in high numbers and subsequently used to produce chimaeric and eventually transgenic animals in which reporter expression is monitored to identify loci which are transcriptionally active during embryogenesis (Gossler *et al.*, 1989; Friedrich and Soriano, 1991).

1.3.1. Enhancer Trapping

The principles underlying enhancer trapping in other systems have been applied successfully in ES cells (Gossler *et al.*, 1989; Korn *et al.*, 1992). The vector p3LSN consisting of a minimal *hsp68* promoter driving *lacZ* and the HSV-tk promoter driving expression of the neomycin phosphotransferase gene (neo^R) was electroporated into ES cells and transformed cell lines selected for using G418. β-gal activity was observed in 30/66 cell lines, either in undifferentiated ES cells or in chimaeric embryos derived from individual cell lines. This indicates that around 45% of enhancer trap integrations are accessing cellular enhancers. Each cell line showing reporter expression in chimaeric embryos 22/66 (33%) gave a unique LacZ expression profile between 7.5 d.p.c. and

10.5 d.p.c. in a range of regional and tissue specific expression patterns (Gossler *et al.*, 1989; Korn *et al.*, 1992).

Enhancer trapping therefore appears an efficient means of accessing cellular enhancers active during embryogenesis. Further analysis of the enhancer trap lines, however, has highlighted limitations in using enhancer trapping as a routine developmental screen in the mouse. Firstly, the mouse genome is not as well characterised as *Drosophila*, making it more difficult to clone the gene associated with the enhancer activity. As a result, endogenous transcripts have been isolated from the vector integration sites from only three enhancer trap lines to date (Korn *et al.*, 1992). Secondly, two of these lines showed more widespread expression of the endogenous transcript when compared to reporter gene activity. This suggests that the vector is incompletely accessing the regulatory elements necessary for the correct expression of the endogenous gene (Soininen *et al.*, 1992; Neuhaus *et al.*, 1994). Finally, no phenotype was observed when two enhancer trap integrations were bred to homozygosity (Korn *et al.*, 1992). Although only two lines were analysed in this study, one would expect relatively few enhancer trap integrations to be mutagenic because entrapment is not dependent on insertion directly into the endogenous gene.

1.3.2. Promoter Trapping

The ability to screen large numbers of ES cells allows for the identification of relatively rare entrapment events. This has made possible the use of promoter trap constructs to overcome the problems inherent in the enhancer trapping strategy. Promoter trap vectors can be divided into three groups; exon, gene, and polyA trap vectors. Exon trap constructs contain a reporter gene without any promoter elements. Reporter activity is dependent on insertion into the exon of an endogenous gene in the correct orientation and reading frame (Figure 1.2A; von Melcher *et al.*, 1992; Friedrich and Soriano, 1991; Macleod *et al.*, 1991). Gene trap vectors differ from exon traps in that they contain a consensus splice acceptor site upstream of the reporter gene.

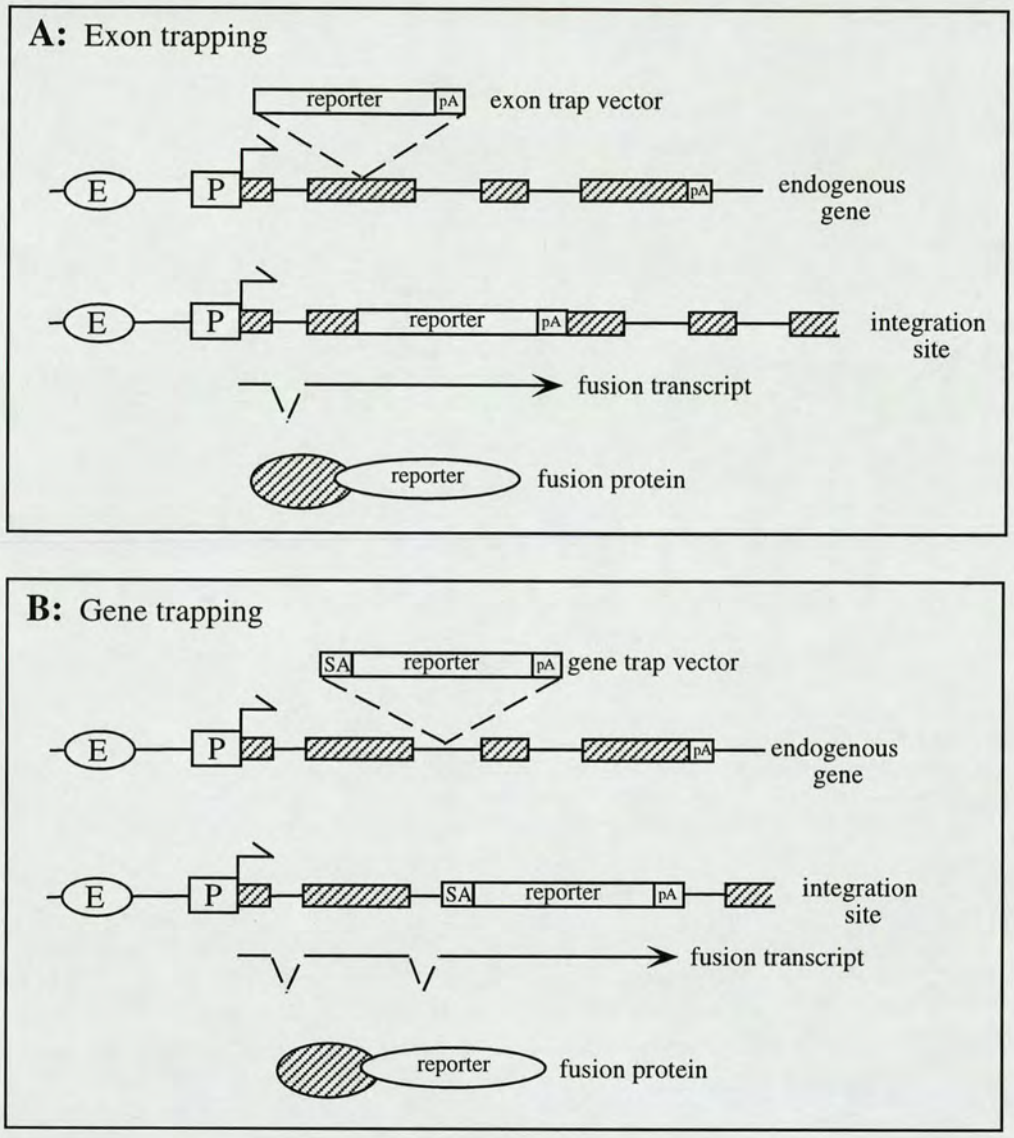


Figure 1.2: Promoter trapping

A: Exon trapping

Exon trap vectors consist only of a reporter gene. Reporter activity is dependent on vector insertion into the exon of a transcription unit in the correct orientation and reading frame. If the reporter has its own translational start codon, vector insertions into untranslated exons can be recovered.

B: Gene trapping

Gene trap vectors contain a splice acceptor sequence upstream of the reporter gene. Vector insertion into the intron of a transcription unit in the correct orientation will result in splicing between the reporter gene and the endogenous upstream exon. Translation of this fusion transcript in the correct reading frame will produce reporter activity. Vector insertions splicing to untranslated exons will be detected if the reporter gene contains its own start codon.

Insertion of the vector, in the correct orientation, into the intron of the endogenous gene will induce splicing from the endogenous upstream splice donor to the vector splice acceptor (Figure 1.2B; Gossler *et al.*, 1989; Friedrich and Soriano, 1991). An active reporter gene will be produced if the reporter is in-frame with the endogenous gene (Skarnes *et al.*, 1992). PolyA trap vectors consist of a reporter gene lacking a polyadenylation (polyA) signal. Insertion of the vector upstream of an endogenous polyA signal produces a functional reporter transcript (Niwa *et al.*, 1993; Figure 1.3; Section 1.3.2.4(v)).

The fact that a promoter trap vector has to integrate into a transcription unit and the subsequent production of a fusion transcript provides three potential benefits for its use in entrapment screens over enhancer traps. Firstly, identification of the endogenous gene should be more straightforward. Secondly, as the reporter is expressed directly from the endogenous promoter, expression should accurately reflect endogenous gene expression. Finally, the integration of the vector is more likely to be mutagenic.
The work reviewed in the next section provides evidence for the above predictions, especially in the context of gene trapping which has been the prevalent system used for ES cell entrapment of developmentally important molecules.

1.3.2.1. Entrapment Aids Gene Identification

The correct insertion of both exon and gene trap vectors induces the production of a transcript consisting of upstream endogenous sequence fused to the reporter gene (Skarnes *et al.*, 1992; Figure 1.2B). This fusion transcript should make the identification of the endogenous gene relatively straightforward using PCR based cloning strategies such as 5'Rapid Amplification of cDNA Ends-PCR (5'RACE-PCR) using primers complementary to vector sequences (Frohman *et al.*, 1988; Skarnes *et al.*, 1992). Refinements to the 5'RACE-PCR protocol by direct sequencing (Townley *et al.*, 1997) has greatly enhanced gene identification and is amenable to automation (Zambrowicz *et al.*, 1998). Endogenous transcripts have also been isolated from

cellular sequences flanking the vector insertion site using inverse PCR, ligation mediated PCR and plasmid rescue protocols (von Melcher *et al.*, 1990,1992; Hicks *et al.*, 1997; Brennan, 1997). The relative ease of gene identification using these techniques has lead to the large scale screening of entrapment events based solely on trapped gene sequence (Zambrowicz *et al.*, 1998; Chowdhury *et al.*, 1997; see Section 1.3.2.5(i)). From these studies and from the individual cell lines characterised (Table 1.1), genes involved in functionally diverse processes are proving susceptible to entrapment (Chowdhury *et al.*, 1997).

1.3.2.2. Reporter Activity Reflects Endogenous Gene Expression

Entrapment screens based on reporter expression during embryogenesis rely on reporter expression correlating well with that of the cellular transcript. From the gene trap lines analysed to date, reporter activity and endogenous transcript expression has been shown to be comparable in most instances (Table 1.1). Increasingly, however, examples where this is not the case are being uncovered (Voss *et al.*, 1998a; Deng and Behringer, 1995; Schuster-Gossler *et al.*, 1998). The transcription factor BTF-3 is expressed ubiquitously throughout the embryo and in the adult. However, a gene trap integration into this gene shows restricted β-gal activity in the heart, limbs and brain (Deng and Behringer, 1995; Table 1.1). Insertion into *Gtl2* provides another example where expression of the endogenous transcript is more widespread than reporter activity (Schuster-Gossler *et al.*, 1994; Table 1.1). Possible explanations for this are vector integrations either into exons alternatively spliced out in certain tissues or affecting enhancer elements altering the transcriptional control of the endogenous gene (Voss *et al.*, 1998a; Deng and Behringer, 1995).

Table 1.1: Characterised Entrapment Events

Vector	Gene	β-gal Activity in Embryo	= Endogenous Transcript*	Null Allele**	Phenotype (defect)	Reference
ROSAβ-geo	TEF-1	widespread	ND	yes	lethal~E11 (heart)	Chen *et al.*, 1994b
"	BTF-3	restricted (heart, limbs, brain)	no	nd	lethal~E6.5	Deng and Behringer, 1995
"	ROSA26	ubiquitous	ND	yes	possibly	Zambrowicz *et al.*, 1997
"	α-enolase	restricted	ND	ND	lethal ~E6.5	Couldrey *et al.*, 1998
"	*dystrophin*	restricted (somites, limbs)	yes	no (h)	muscle degeneration	Wertz and Fuchtbauer, 1998
U3β-geo	ECK	restricted (ps, node, hindbrain)	yes	yes	none	Chen *et al.*, 1996
"	ArMT	restricted (brain, neural tube)	ND	ND	embryonic lethal	Scherer *et al.*, 1996
U3neo	fug1	widespread	NA	yes	lethal~E6.0	DeGregori *et al.*, 1994
U3tkneo	XVIII-1	NA	NA	yes	viable	Muth *et al.*, 1998
pGT4.5	Gt2	widespread	ND	ND	viable	Skarnes *et al.*, 1992
"	Gt4-1	restricted (neural)	ND	yes	perinatal lethal	"
"	Gt4-2	widespread	yes	yes	growth retarded/ lethal	"
"	Gt10	restricted (otocyst, bv, gut, bladder)	yes	yes	ND	"
"	*cordon bleu*	restricted (node, fp, notochord, gut, liver)	yes	no (r)	viable	Gasca *et al.*, 1995
PT1-ATG	I.23	restricted (dorsal midbrain)	ND	ND	viable	Forrester *et al.*, 1996
"	I.114	restricted (liver)	ND	ND	viable	"
"	I.163	restricted (ys, somites, spinal cord)	ND	ND	viable	"
"	R.140	restricted (lb)	ND	ND	embryonic lethal	"
"	*aquarius*	restricted (lb, somites, branchial arches)	yes	no	viable	Sam *et al.*, 1998
"	TFEB	restricted (ys, heart, liver)	yes	no	viable	McClive *et al.*, 1998
"	R.108	resricted (heart, neural tube, lb)	ND	ND	viable	"
"	R.124	restricted (heart)	ND	ND	postnatal lethal (heart)	"
pGT1.8β-geo	E-catenin	restricted (epithelia)	yes	yes	lethal~E3.5-4.0 (trophoblast)	Torres *et al.*, 1997
"	MAP-4	widespread	yes	no	viable	Voss *et al.*, 1998b
"	PTP-BL	restricted (epithelia, PNS)	yes	ND	viable	Thomas *et al.*, 1998b
SA-IRESβ-geo	*bodenin*	restricted (head, heart)	yes §	no	viable	Faisst and Gruss, 1998
"	*paddy*	restricted (lb)	yes §	ND	viable	Pires-DaSilva and Gruss, 1998
pGT1.8TM	LAR	widepread	yes	yes	viable	Skarnes *et al.*, 1995
"	PTPκ	restricted (endoderm, pm, somites)	yes	yes	viable	"
"	*netrin*	restricted (spinal cord, fp, somites)	yes	no (h)	postnatal lethal	Serafini *et al.*, 1996
"	HS2ST	restricted	yes	yes	perinatal lethal (renal agenesis)	Bullock *et al.*, 1998
TV2	*jumonji*	restricted (mid-hindbrain, cerebellum)	ND	yes	lethal~E15.5 (neural tube)	Yoshida *et al.*, 1995
pGTi	*Gtl2*	restricted (myogenic lineage)	no	no (h)	dwarfism	Schuster-Gossler *et al.*, 1996+1998
pPAT	PAT12	none	no	yes	viable	Yoshida *et al.*, 1995

SEE ALSO: Stoykova *et al.* (1998)

ND. analysis not done. NA. not applicable.
*, does the expression of the reporter gene accurately reflect the expression pattern of the endogenous gene?
**, judged by the absence of endogenous gene sequence 3' of vector insertion site in animals homozygous for the entrapment event.
§, very limited examination of reporter versus endogenous transcript expression; r, reduced level of endogenous transcript; h, hypomorphic allele
ps, primitive streak; bv, blood vessels; fp. floor plate; ys, yolk sac; lb, limb bud;

1.3.2.3. Vector Insertion Is Mutagenic

The mutagenicity of exon and gene trap constructs relies on the production of a fusion transcript terminating at the reporter polyA signal effectively ablating the cellular transcript downstream of the insertion site (Figure 1.2).

Data from exon and gene trap screens carried out to date indicates that vector integration produces recessive lethal phenotypes in 39/104 (37%) of lines examined (Camus *et al.*, 1996; Skarnes *et al.*, 1992; Hicks *et al.*, 1997, Friedrich and Soriano, 1991; Forrester *et al.*, 1996; Stoykova *et al.*, 1998; Voss *et al.*, 1998a) correlating well with the percentage of phenotypically lethal loci identified during random mutational screens in other model systems (Miklos and Rubin, 1996).

More directly, the detailed characterisation of promoter trap lines confirms that in the majority of cases, vector integration disrupts the endogenous transcript 3' of the vector integration site (Table 1.1). In four lines vector integration has resulted in reduced but detectable levels of endogenous transcript in animals homozygous for the promoter trap integration. Gene trap integrations into *netrin* (Serafini *et al.*, 1996), *dystrophin* (Wertz and Fuchtbauer, 1998) and *Gtl2* (Schuster-Gossler *et al.*,1998) result in hypomorphic alleles of these genes, whereas insertion into the fourth gene, *cordon bleu* (Gasca *et al.*, 1995) reveals no obvious phenotype (Table1.1). It remains to be seen whether a lack of phenotype in the last line is due to sufficient production of endogenous message or redundant gene function. Most interestingly, in a further four gene trap lines, *TFEB* (McClive *et al.*,1998), *bodenin* (Faisst and Gruss, 1998), *aquarius* (Sam *et al.*, 1998), and MAP-4 (Voss *et al.*, 1998b) endogenous expression levels approaching wild type are observed in animals homozygous for the gene trap integration. These lines are thought to reflect instances where mRNA processing splices around the integrated vector due to inefficient use of the vector polyA signal (Sam *et al.*, 1998; Voss *et al.*, 1998b).

Promoter trapping in ES cells therefore provides a practical, single route to identify and mutate developmentally significant molecules in the mouse (Table1.1).

Insertional mutants have been produced, affecting a range of developmental processes including heart formation (Chen *et al.*, 1994b), neural tube closure (Takeuchi *et al.*, 1995) and axon guidance (Serafini *et al.*, 1996). Promoter trap integrations are also producing excellent tissue specific and cell lineage markers with which to further study development. Markers specific for cardiac tissue (McClive *et al.*, 1998) and neuronal cells (Stoykova *et al.*, 1998) are just two examples of studies producing potentially useful cell lines. One particularly well characterised gene trap line, ROSA-βgeo 26, displays reporter activity in all cell types examined making it an excellent marker for chimaera experiments such as haematopoetic cell transplantation studies (Zambrowicz *et al.*, 1997).

1.3.2.4. Promoter Trapping Technology

The data outlined above, highlighting the potential of promoter trapping, has been achieved using widely varying entrapment technologies. This next section reviews the different vector designs and vector delivery systems which have been developed to increase the efficiency of, and eliminate any biases in, promoter trapping in ES cells.

i/ Vector design

The application of promoter trap technology to the mouse, both *in vitro* or *in vivo*, has spawned many variations on the basic promoter trap structure. The inclusion of the translational start site for the reporter gene allows for the recovery of promoter trap insertions into untranslated regions of genes. Subsequently, an improvement in the efficiency of these promoter trap vectors to access true entrapment events is observed, that is, an increase in the proportion of neomycin expressing (G418 resistant) transformants displaying β-gal activity (Brenner *et al.*, 1989; Hill and Wurst, 1993).

In the production of gene trap vectors, the inclusion of the splice acceptor site upstream of the *lacZ* gene was shown to increase gene trap efficiency when compared directly to exon trap vectors (Gossler *et al.*, 1989; Friedrich and Soriano, 1991). This greater efficiency reflects the larger genomic target that introns provide for vector insertions compared to the smaller potential sites available to exon trap vectors (Friedrich and Soriano, 1991). It is of interest to note that gene trap vectors inserting into exons in frame with the reporter gene will also produce a functional reporter protein using cryptic splice donor sites within the vector intron sequence (McClive *et al.*, 1998; Brennan, 1997).

A feature in several gene trap vectors is the inclusion of a viral internal ribosomal entry site (IRES) between the splice acceptor and reporter gene (Chowdhury *et al.*, 1997; Takeuchi *et al.*, 1995; Zambrowicz *et al.*, 1998). The presence of the IRES is predicted to produce the independent translation of the endogenous gene and the reporter gene from a single dicistronic message (Mountford and Smith, 1995). Chowdhury *et al.* (1997) showed that the inclusion of an IRES increased both the number of G418 resistant transformants and proportion of these transformants with β-gal activity. In using an IRES, reporter expression will be independent of the endogenous reading frame into which the vector has inserted. Furthermore, any potential steric hinderance of reporter or selection activity by endogenous N-terminal protein fusion should be eliminated. The design of the secretory trap vector pGT1.8TM features the CD4 transmembrane domain between the splice acceptor and reporter gene, preselecting for the entrapment of genes producing secreted or membrane bound proteins (Skarnes *et al.*, 1995). This vector is reviewed as a preselection strategy in Section 1.3.2.5(ii).

ii/ Reporter gene

Several different reporter genes have been used for trapping in ES cells. The drug resistance genes histidine dehydrogenase and neomycin phosphotransferase have

been used, providing excellent selection *in vitro* (von Melcher *et al.*, 1990, 1992). However, large N-terminal fusions may compromise the activity of these reporters (Macleod *et al.*, 1991) and activity cannot be assayed for, temporally or spatially. Accordingly, these reporters, in particular neomycin, have been used in more general entrapment screens (von Melcher *et al.*, 1992; Hicks *et al.*, 1997).

In gene trap screens designed to identify developmentally regulated genes, a readily assayable reporter is a prerequisite. Consequently, promoter driven neomycin expression has been used as selection for transformants with the subsequent selection of entrapment events relying on the expression of the *lacZ* reporter gene (Gossler *et al.*, 1989; Friedrich and Soriano, 1991). As alluded to earlier, β-gal activity is easily assayed for in ES cells as well as embryos and can tolerate large N-terminal fusions (Friedrich and Soriano, 1991; Casabadan, 1980). Nevertheless, in having promoter driven selection independent of reporter expression it is necessary to screen out integrations outwith transcription units.

The β-geo reporter gene, consisting of neomycin fused to the 3' end of β-galactosidase, was developed to eliminate this allowing for the direct selection of promoter trap events in ES cells (Friedrich and Soriano, 1991). Theoretically, all neomycin expressing clones will have β-galactosidase activity. However, in practice, the greater sensitivity of G418 selection results in entrapment events into genes whose expression level is sufficient to confer drug resistance but below the threshold necessary for X-gal staining (Friedrich and Soriano, 1992; Skarnes *et al.*,1995). Depending on which allele of neomycin is used (Yenofsky *et al.*, 1990), 95% of gene trap integrations show reporter activity using the hypoactive allele (Friedrich and Soriano, 1992) with 60% and 30% observed using the wild type allele in the vectors pGT1.8β-geo (Skarnes *et al.*,1995) and pSAβ-geo (Chowdhury *et al.*, 1997) respectively. In using the hypoactive allele, a preselection is observed for cell lines with high levels of β-gal activity mirroring the high level of *neomycin* expression needed for G418 resistance. By contrast, use of the wild type allele resulted in more varied levels of β-gal activity which highlights the greater difference in sensitivity

between β-gal activity and wild type neomycin (Skarnes *et al.*, 1995). This allows for the trapping of genes expressed at lower levels and increases the potential number of genes accessible to entrapment when compared to the hypoactive allele (Skarnes *et al.*, 1995).

Alkaline phophatase (AP) has also been used as a reporter gene in exon and gene trap vectors as an alternative to β-gal activity for the monitoring of expression during development (Xiong *et al.*, 1998). The efficiency of entrapment using the AP reporter was comparable to using the *lacZ* reporter gene. Moreover, AP activity could be monitored during *in vitro* differentiation prototcols and *in vivo* providing excellent spatio-temporal reporting of endogenous gene expression (Xiong *et al.*, 1998).

It is of interest to note recent work in the development of green flourescent protein (GFP) as a reporter gene in mammalian systems. GFP has been used successfully for entrapment protocols in slime moulds (Fey and Cox, 1997) and benefits over other reporters in that its expression can be monitored non-destructively in living cells (Heim *et al.*, 1994). In the mouse, GFP is efficiently detected *in vivo* and *in vitro* and can be transmitted through the germline (Hadjantonakis *et al.*, 1998). Moreover, the detection of *GFP-Hox* gene fusions during mouse embryogenesis has recently been reported by Godwin *et al.* (1998). These studies provide the foundation for the application of GFP technology to produce more refined and powerful entrapment screens in the mouse.

iii/ Vector delivery

Entrapment vectors have been introduced into ES cells either via electroporation or retroviral infection (Gossler *et al.*, 1989; Friedrich and Soriano, 1991; von Melcher *et al.*, 1992). The use of either technique can have a direct bearing on the resulting vector integrations.

Electroporation of entrapment vectors is an attractive procedure due to its relative simplicity, the relative inefficiency of which can be overcome by using larger

numbers of ES cells (Skarnes *et al.*, 1992). However, the electroporation of entrapment vectors has been shown to induce insertional artefacts (Friedrich and Soriano, 1991; Niwa *et al.*, 1993; Takeuchi *et al.*, 1995; Skarnes *et al.*, 1992). The deletion of extreme sequences from plasmid entrapment vectors has been observed in a number of lines (Niwa *et al.*, 1993; Takeuchi *et al.*, 1995; Torres *et al.*, 1997). Despite this, vector function is generally uncompromised as the deletion affects non-essential external sequences.

From the limited number of plasmid vector integration sites examined, two integrations induced uncharacterised rearrangements in the cellular DNA (Niwa *et al.*, 1993) while another three insertions caused no obvious effects (Soininen *et al.*, 1992; Macleod *et al.*, 1991). How common these rearrangements are is unknown but rearrangements perturbing the transcriptional control of the endogenous gene are relatively rare because in the majority of plasmid integrations analysed to date reporter expression correlates well with the endogenous gene (Table 1.1).

Another characteristic of electroporation is the integration of mutiple copies of the plasmid vector (Friedrich and Soriano, 1991; Forrester *et al.*, 1996; Takeuchi *et al.*, 1995; Skarnes *et al.*, 1992). The majority of multiple vector insertions characterised appear to be in concatemeric arrays ranging from 2-5 vector copies at a single chromosomal site (Friedrich and Soriano, 1991; Skarnes *et al.*, 1992; Takeuchi *et al.*, 1995). However, backcrossing mice from three independently derived gene trap lines segregated individual vector copies suggesting that vectors had inserted at 2 distinct chromosomal loci in these lines (Forrester *et al.*, 1996, Voss *et al.*, 1998a). As with deletion of vector sequences, integration of multiple vectors is predicted not to interfere with entrapment function as only the vector at the extreme of the tandem repeat should be active. Indeed in the majority of cases examined to date single fusion transcripts are produced, the vector is mutagenic and reporter activity mimics the endogenous gene (Forrester *et al.*, 1996; Takeuchi *et al.*, 1995; Skarnes *et al.*, 1992). Nevertheless, it is noteworthy that reporter expression level from concatemeric transgene arrays is

negatively affected by an increased number of transgene copies integrated (Garrick *et al.*, 1998; Wolffe, 1998).

Retroviral entrapment vectors typically contain the reporter gene within the body of the virus if it has a splice acceptor sequence upstream (Friedrich and Soriano, 1991; Brenner *et al.*, 1989) or within the U3 region of the LTR if the reporter lacks a splice acceptor (von Melcher *et al.*, 1992). The mechanism of retroviral infection of ES cells is very precise so that at low multiplicities of infection a single proviral entrapment vector integrates into the genome while the structure of the vector and cellular DNA is preserved (von Melcher *et al.*, 1992; Chen *et al.*, 1996; Scherer *et al.*, 1996).

Retroviruses and their entrapment derivatives have preferential intragenic and possibly intergenic insertion sites (Jaenisch, 1988). Retroviruses have been shown to integrate close to DNaseI hypersensitive sites (Vijaya *et al.*, 1986, Rhodewold *et al.*, 1987) and transcriptionally active regions of the genome (Mooslehner *et al.*, 1990; Scherdin *et al.*, 1990). Moreover, retroviral entrapment vectors preferentially integrate close to the 5' end of transcription units as judged by the length of the resulting fusion transcript and promoter activity associated with genomic regions immediately 5' of the insertion site (von Melcher *et al.*, 1990; Friedrich and Soriano, 1991). Consequently, 5' retroviral insertions are more likely to interfere with *cis*-acting transcriptional control elements as observed in the transgenic retroviral line Mov13 (Hartung *et al.*, 1986; Barker *et al.*, 1991). As alluded to previously, this may be the case with the retroviral gene trap insertion into the transcription factor BTF-3 where reporter expression does not correlate with the endogenous transcript (Deng and Behringer, 1995). Interestingly, a study of integrations into known genes suggests that no such bias occurs with plasmid vectors (Chowdhury *et al.*, 1997).

On a genome wide scale, studies on avian retroviruses reveals that insertions occur at specific genomic sites at a very high frequency (Shih *et al.*, 1988; Withers-Ward *et al.*, 1994). However, this doesn't exclude other sites from insertion as most regions of the genome are predicted to be accessible to retroviral integration at varying frequencies (Withers-Ward *et al.*, 1994).

iv/ Entrapment is a random event

An important issue raised by retroviral integration is whether there are any potential biases in the insertion of entrapment vectors. As larger, saturation level entrapment screens are undertaken, it is important to know the number of gene loci accessible to entrapment and whether certain genes are resistant or more susceptible to entrapment.

In *Drosophila* , an entrapment screen designed to isolate nervous system phenotypes, identified repeated P-element insertions within individual genes (Bier *et al.*, 1989). Moreover, certain chromosomal regions appear to be resistant to P-element insertion (Smith *et al.*, 1993).

In ES cells, the analysis of promoter trap insertions into 400 loci revealed 3 insertions each into L29 and α-NAC (Hicks *et al.*, 1997) and from 6 secretory trap lines, integration had occurred into two distinct sites of the LAR gene (Skarnes *et al.*, 1995). Moreover, entrapment events have been repeated by different vectors. *Jumonji* was initially identified by Takeuchi *et al.* (1995) using the TV2 gene trap vector and has since been trapped using the retroviral ROSAβ-geo and plasmid pGT1.8TM and pGT1.8βgeo gene trap vectors (Baker *et al.*, 1997; Voss *et al.*, 1998a; W. C. Skarnes, personal communication). Similarly, insertions into both enolase and R-PTPκ have been reported, again by different vectors (Chowdhury *et al.*, 1997; Skarnes *et al.*, 1995; Couldrey *et al.*, 1998). Even with the relatively small number of integrations analysed to date it is apparent that certain genes are more frequently identified during entrapment. The precise nature of this bias in unknown but may reflect a number of different factors. Genes comprising of large intronic regions are predicted to be more accessible to gene trap vectors (Skarnes *et al.*, 1995). A more open chromatin structure associated with transcriptionally active loci is known to influence retroviral integration (Jaenisch, 1988) and could potentially affect plasmid vector insertion. It could also reflect the high expression level of the genes facilitating their identification by different entrapment protocols. Despite this, a small percentage of over-represented loci is not

predicted to interfere significantly with the number of total loci accessible to entrapment (Evans, 1998).

v/ PolyA trapping

A significant limitation of exon and gene trapping is that identification of true entrapment events is reliant on reporter activity which in turn is reliant on expression of the trapped endogenous gene in ES cells. Given that certain developmentally important genes are not expressed in ES cells, for example MyoD and Myf5 (Rohwedel *et al.*, 1994), exon and gene trapping will fail to access such genes. PolyA trap vectors overcome these limitations. The first polyA trap vectors to be developed contained a phosphoglycerate kinase (PGK) promoter-neomycin cassette lacking a polyA addition signal (Niwa *et al.*, 1993). In ES cells, promoter driven neomycin is dependent on a downstream endogenous polyA signal to be stably expressed and is therefore independent of the expression status of the endogenous gene. In addition, a *lacZ* reporter fused to a consensus splice acceptor site was located upstream of PGK-neomycin allowing for endogenous gene expression to be monitored (Niwa *et al.*, 1993). However, inefficient vector splicing caused by deletions of vector sequences resulted in the preference for vector integrations into the 5' end of genes. The subsequent difficulties in isolating definitive polyA addition signals using 3' RACE-PCR made it difficult to appraise the effectiveness of this polyA trap vector (Niwa *et al.*, 1993).

A more robust polyA trap vector (pPAT) was developed by Yoshida *et al.* (1995). This vector has two modifications; intronic sequence upstream of the splice acceptor site to insulate the splicing activity from vector deletions and a splice donor sequence downstream of the neomycin gene for more efficient polyA trapping (Figure 1.3; Yoshida *et al.*, 1995). Using pPAT, 49 G418 resistant lines (polyA trap events) were produced of which 12 (25%) gave amplification products after 3'RACE, substantiating that vector insertions are more common towards the 5' end of genes

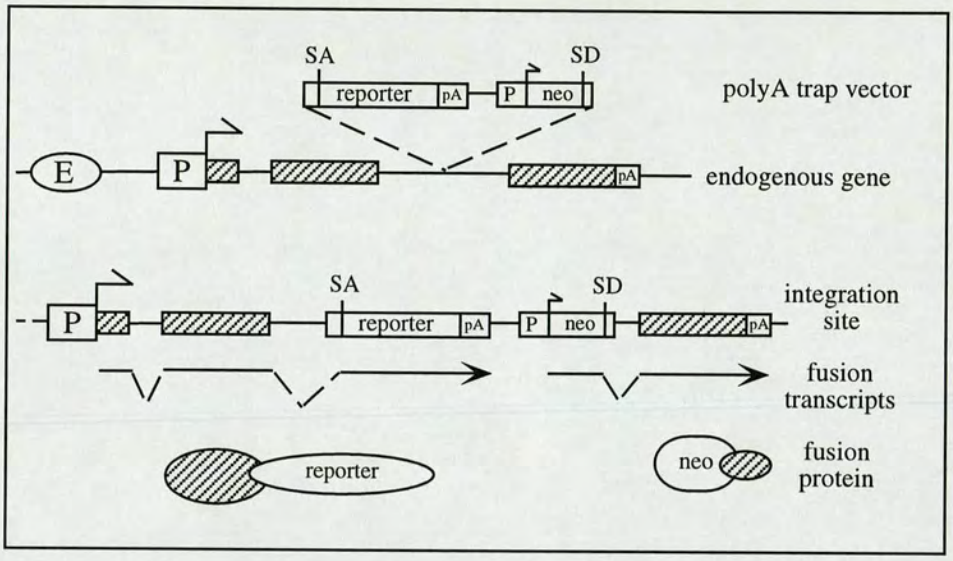


Figure 1.3: PolyA trapping

PolyA trap vectors contains a promoter driven selectable marker (*neo*) lacking a polyadenylation signal downstream of a reporter gene. The recovery of vector insertion into a transcription unit is dependent on the selectable marker acquiring an endogenous polyA signal via the splice donor site (SD). The upstream reporter gene acts as a gene trap and allows for the monitoring of endogenous gene expression.
SA, splice acceptor; pA, polyadenylation signal

which are less likely to be amplified using 3'RACE. From 6 of these amplifiable cell lines, 5 contained polyA addition signals, 4 of which were from novel transcripts. Furthermore, β-galactosidase activity was observed in chimaeric embryos derived from 3 of the 4 novel sequence producing lines. The vector was also shown to be mutagenic, disrupting the endogenous transcript of a single line (Yoshida *et al.*, 1995).

Recently, polyA trapping in ES cells has been used to produce a library of sequence tags from 2000 genes potentially mutated by vector insertion (Zambrowiscz *et al.*, 1998; see Section 1.3.2.5(i)). PolyA trapping therefore provides an efficient method of identifying and mutating genes independently of expression state in ES cells.

vi/ Gene trapping and site specific recombination

A novel promoter trap strategy involving site-specific recombination was initially developed using an IL-3 dependent haematopoietic cell line to identify genes activated during programmed cell death (Russ *et al.*, 1996). A similar strategy has since been applied to identify transcripts transiently induced during development (Thorey *et al.*, 1998). A stable ES cell line was produced containing a reporter cassette comprising a PGK promoter driving a *loxP* flanked neomycin cassette upstream of a *lacZ* reporter (Figure 1.4). A novel gene trap vector comprising Cre recombinase was introduced into this cell line. Insertion of the gene trap vector into an active gene produces Cre activity, excising neomycin from the reporter cassette, resulting in a permanent switch to LacZ expression (Figure 1.4). Using this strategy a total of 5 gene trap cell lines not expressing Cre (neomycin expressing) were differentiated *in vitro*, with two of these lines subsequently used to produce transgenic mice. In both differentiating ES cells and transgenic embryos, β-gal activity was observed indicating that initial neomycin expression had switched to lacZ via Cre recombination activated from expressed cellular transcripts. Confirming the ability of this system to detect transiently induced transcripts, the endogenous genes from two lines with β-gal activity

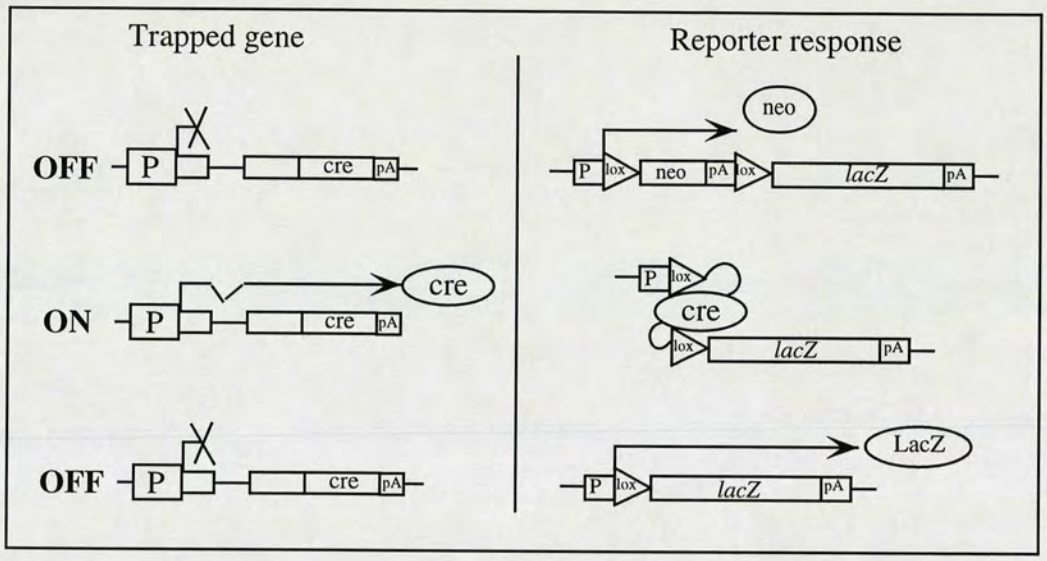


Figure 1.4: Gene trapping using site specific recombination

The Cre recombinase gene is introduced into a transgenic ES cell line containing a reporter cassette with promoter driven neomycin expression upstream of a silent *lacZ* gene. Insertion of the Cre gene into a endogenous transcription unit has no effect on neomycin expression when the endogenous gene is inactive. Activation of the endogenous gene produces Cre under the control of the promoter of the endogenous gene. Cre mediated excision of the neomycin gene via flanking loxP sites places the *lacZ* gene under the control of the neomycin promoter. Expression of *lacZ* is maintained even if the endogenous locus stops expressing Cre.
(Adapted from Thorey *et al.*, 1998)

were shown to be only transiently induced upon ES cell differentiation (Thorey *et al.*, 1998).

Consequently, conditional entrapment provides a complementary system to promoter trapping, as entrapment using this strategy does not rely on expression of the trapped gene in ES cells. Indeed genes expressed in ES cells are selected against. Furthermore, the permanent switch in reporter expression by Cre activation can be induced by genes only transiently expressed during development although the spatial and temporal expression of individual genes will be impossible to monitor.

1.3.2.5. Prescreening Gene Trap Events

With potentially every gene in mice accessible to entrapment by existing vectors (Hicks *et al.*, 1997; Gossler *et al.*, 1989; Friedrich and Soriano, 1991, von Melcher *et al.*, 1992), the criteria by which cell lines are selected for further study depends on the individual investigators interests. Using ES cells, however, allows for the pre-selection of cell lines before attempting germ line transmission of the integration for phenotype analysis. This section examines techniques which have been used to pre-screen gene trap cell lines. Comparative analysis of endogenous gene sequence, biases introduced by vector design and monitoring reporter gene expression *in vivo* or during *in vitro* differentiation protocols have been used to pre-select integration events for further analysis.

i/ Sequence

As mentioned previously in Section 1.3.2.1, improvements in identifying endogenous gene sequence have made possible the large scale screening of integration events. Comparative sequence analysis to cDNA databases of sequences from three such screens is shown in Table 1.2. Promoter proximal sequence tags (PST's) represent upstream genomic sequences from retroviral vector insertion sites and, as

Table 1.2: Sequence data from 3 large scale entrapment screens

VECTOR	TARGET/ SEQUENCE DATA	TOTAL	HOMOLOGY (%total)		
Plasmid gene-trap			known genes	EST's	novel
pSAβgeo/pSAIRESβgeo (1)	exon/5'RACE-PCR	56	17(30%)¥	11(20%)	28(50%)
pGT1.8TM (2)	exon/Direct Sequence	57	29(51%)¥	11(19%)	17(30%)
Retroviral Promoter Trap					
U3neo SV1 (3)	PST*/plasmid rescue	400	42(11%)	21(5%)	337(84%)

(1) Chowdhury *et al.*, 1997
(2) Townley *et al.*, 1997
(3) Hicks *et al.*, 1997
*PST= Promoter Proximal sequence Tag
¥ - The larger proportion of known genes identified using pGT1.8TM reflects the limited type of genes (secretory/membrane spanning) this vector can access (Skarnes *et al.*, 1995).

such, potentially include promoter or intronic sequences because a retroviral promoter trap event is not dependent on insertion into an exon (Hicks *et al.*, 1997; von Melcher *et al.*, 1992). RACE-PCR based techniques identify exon sequences only and therefore gives a more accurate prediction of the proportion of known genes, ESTs and novel genes accessed in these screens (Chowdhury *et al.*,1997). Pre-screening using gene sequence allows for the selection of integrations into genes with, for example, structural homology to important developmental genes or candidate disease genes. It also allows for the exclusion of previously characterised genes from further study.

Perhaps more significantly, given the massive amounts of EST sequence data available (Boguski *et al.*, 1993), these screens provide a valuable resource for attributing gene sequence data to biological function on a large scale. This is highlighted by the recent characterisation of sequence from over 2000 gene trap insertions (Zambrowicz *et al.*, 1998). The systematic application of these techniques has the potential to produce mutations in every gene expressed in the mouse.

ii/ Secretory trap vectors

The secretory trap vector was developed as a means of identifying the secreted and transmembrane proteins important for cell signalling in the developing embryo (Skarnes *et al.*, 1995). The secretory trap vector (pGT1.8TM) contains the CD4 transmembrane (TM) domain between the splice acceptor and β-geo. Insertion of the vector immediately downstream of an endogenous signal sequence or downstream of a signal sequence and an even number of transmembrane domains is predicted to place βgeo in a typeI orientation, maintaining reporter activity in the cytosol (Figure 1.5). Integration of the secretory trap vector into a gene either lacking a signal sequence or downstream of a signal sequence and an odd number of transmembrane domains places β-geo into the endoplasmic reticulum (ER) lumen abolishing activity (Figure 1.5). Data from six secretory trap lines characterised to date shows that vector insertion into

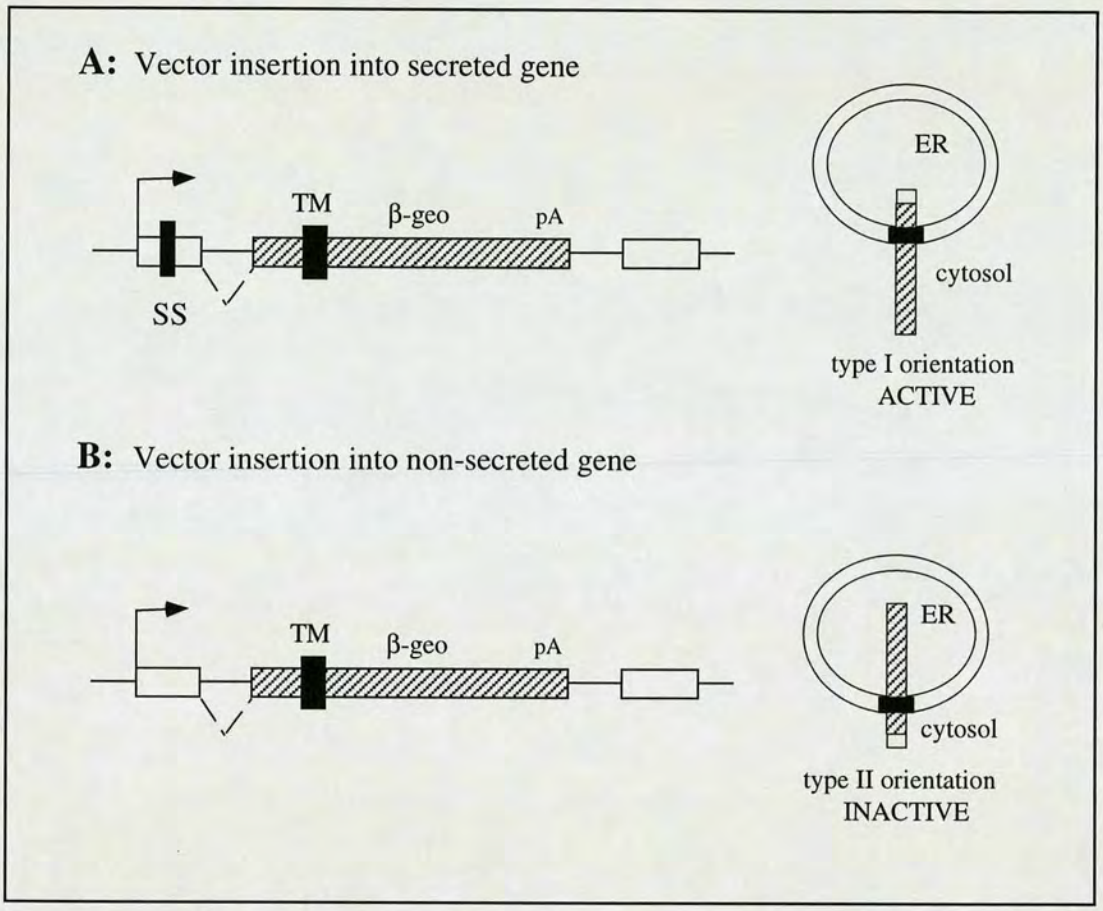


Figure 1.5: Model of secretory trap vector action

Diagram showing the basis of secretory trap selection for vector insertions downstream of signal sequences

A: Insertion of the secretory trap vector into genes containing a signal sequence (SS) produces a fusion protein that inserts into the endoplasmic reticulum (ER) membrane in a type I configuration. The vector transmembrane domain (TM) places the β-geo reporter in the cytosol where β-gal is active.

B: Insertion of the secretory trap vector into a gene lacking a signal sequence produces a fusion protein which inserts in the ER membrane in a type II orientation. Consequently, β-geo is exposed to lumen of the ER where β-gal is inactive.

(Figure adapted from Skarnes *et al.*, 1995)

membrane spanning molecules has occurred downstream of the endogenous signal sequences as predicted (Skarnes *et al.*, 1995).

iii/ Reporter expression

The main driving force in gene entrapment technology thus far has been the potential to isolate novel genes with spatially and temporally restricted expression patterns during embryogenesis (Gossler *et al.*, 1989). To this end Wurst *et al.* (1995) analysed reporter gene activity in chimaeric embryos derived from 279 independent gene trap cell lines. Chimaeric embryos were initially dissected at 8.5 d.p.c. From 279 cell lines, 36 (13%) showed spatial variability in reporter activity, 88 (32%) produced widespread or ubiquitous activity and 155 (55%) gave no reporter gene activity. A quarter of the cell lines examined at 8.5 d.p.c. were also examined at 12.5 d.p.c. for reporter activity. Of the cell lines showing ubiquitous or no reporter activity, around 30% showed alternative patterns of reporter activity at 12.5 d.p.c. Extrapolating these data infers that approximately 40% of gene trap integrations isolated in ES cells exhibit temporal or spatial variability in reporter activity during embryogenesis. In another study, Voss *et al.* (1998a) report slightly different proportions of chimaeric embryos showing ubiquitous (30%), restricted (32%) and no (38%) reporter activity in a study of 183 cell lines at E11.5. The application of tetraploid aggregation chimaera technology, which allows for the almost complete derivation of embryos from ES cells (Nagy *et al.*, 1990), would reduce the number of chimaeric embryos needed to be confident of an complete reporter gene expression profile, saving on time and animals.

Despite the obvious benefits in screening gene trap lines in this manner, the routine production of chimaeras from each cell line is not a realistic proposition in terms of time and cost when pre-screening large numbers of cell lines for developmentally important genes.

iv/ *In vitro* prescreening

In order to enrich for gene trap events into developmentally regulated genes prior to chimaera production, several groups have studied the reporter gene activity of individual gene trap cell lines during *in vitro* differentiation protocols.

Baker *et al.* (1997) examined reporter gene activity in embryoid bodies (EB) derived from 86 ROSAβgeo cell lines. EB are aggregates of differentiated ES cells which are capable of producing various embryonic cell types prevalent in the early stages of embryonic development. Blood islands and blood vessels, cardiomyocytes, skeletal myocytes, chondrocytes, neurons and glia are produced spontaneously in EB. Analysis of the 86 gene trap lines revealed that 22% showed no reporter activity in EB, 23% had ubiquitous activity and 55% gave varying degrees of restricted reporter activity. Co-expression of β-galactosidase and lineage specific markers for skeletal myoblasts, cardiomyocytes, chondrocytes and neurons was observed in the subset of cell lines displaying restricted reporter expression. Furthermore, a good correlation was observed in these lines between *in vitro* reporter activity and *in vivo* expression of the trapped gene. This correlation was also observed in a study of 191 U3β-geo lines differentiated to form EB. From this screen, 16% showed modulated reporter activity in EB of which seven lines transmitted through the germline displayed comparable alterations in reporter activity in the embryo (Scherer *et al.*, 1996). More recently, two similar reports highlight the fact that even very limited *in vitro* pre-selection protocols can achieve a two-fold enrichment for genes with restricted expression during development by reducing the proportion of lines selected which result in ubiquitous reporter expression in the embryo (Voss *et al.*, 1998a; Stoykova *et al.*, 1998).

Another strategy to enrich for developmentally expressed genes involves the pre-selection of gene trap integrations in which reporter activity was modulated by the differentiating factor retinoic acid (RA) (Forrester *et al.*, 1996). Variations in the physiological levels of RA in the embryo induces a range of teratogenic effects (Linney, 1992, Eichele 1989) which are recapitulated in embryos deficient in individual or

groups of retinoic acid receptors (Kastner *et al.*, 1997). Moreover, the RA signalling pathway has been shown to modulate the expression of a number of developmentally important genes *in vitro* and *in vivo* (Pruitt, 1992; Simeone *et al.*, 1990; Conlon and Rossant, 1992; Marshall *et al.*, 1992).

From 202 β-gal positive cell lines treated with RA, reporter activity was induced in 9 lines and repressed in 11. Tetraploid aggregation chimaeras derived from these 20 lines showed spatially restricted reporter activity in all but one line from 8.5 -11.5 d.p.c./late gastrulation to midgestation. Of the 11 repressed lines 8 showed reporter activity in the heart, 4 of which also displayed activity in the hindbrain, craniofacial region and branchial arches. From the 9 induced lines, 3 showed reporter activity in the region of the spinal cord and adjacent somites. These results, therefore, show a significant enrichment for developmentally restricted genes not seen using other protocols. A similar enrichment is seen when ES cells are differentiated into EB in the presence of RA (Gavojic *et al.*, 1998). Interestingly, the majority of tissues with reporter activity in these embryos also display patterning defects induced by variations in RA levels suggesting that the trapped genes may be downstream of the RA signalling pathway (Forrester *et al.*, 1996). Although this will ultimately limit the number of integrations into genes that are responsive to RA, this system can be applied to isolate genes modulated by other differentiating factors. Bonaldo *et al.* (1998) differentiated gene trap cell lines in the presence of the growth factors follistatin and nerve growth factor as well as RA. As with the previous screens, a modulation in reporter activity *in vitro* correlated well with developmentally restricted activity *in vivo*.

1.3.2.6. Summary

Promoter trapping in ES cells therefore provides a powerful means for the identification and functional analysis of individual genes. Random large scale entrapment mutagenesis programmes catalogued by gene sequence data as carried out by Zambrowiscz *et al.* (1998) complement existing expressed sequence databases and

genome sequencing projects and perhaps provide the most realistic means of achieving saturation level mutagenesis in the mouse. Lexicon Genetics provides these sequences and cell lines through Omnibank™ as a commercial resource to investigators for the functional analysis of the trapped genes. The large scale functional screening of individual mutants from such a resource is financially and ethically prohibitive. The pre-screening of entrapment events provides a solution to this as it allows for more focused searches for genes of interest based on expression pattern or gene sequence structure (Forrester *et al.*, 1996; Skarnes *et al.*, 1995).

Using retinoic acid pre-screening, the I114 gene trap cell line was identified which showed restricted reporter activity in the liver of mid-gestation chimaeric embryos (Forrester *et al.*, 1996). This thesis reports the characterisation of the I114 cell line.

1.4. LIVER DEVELOPMENT

In contrast to the developing mesoderm and ectoderm, relatively little is known of the molecules mediating the patterning of the endodermal lineage. The liver is the earliest derivative of the gut endoderm providing an attractive system with which to investigate endodermal differentiation. The anatomy of liver development is relatively well characterised involving the interaction of a number of tissues, the nature of which parallel organogenesis in other tissues. Moreover, recent molecular analysis has highlighted potentially conserved mechanisms of gut endoderm differentiation between the mouse and other model systems such as *Drosophila* and *C.elegans*. Insights gleaned from the study of liver development may identify factors with functional overlap in hepatic regeneration and disease states such as cancer, as has already been seen for certain foetal markers.

1.4.1. Overview Of Liver Development

In the mouse, liver development commences at around 8.5 d.p.c. when the ventral floor of the foregut endoderm thickens forming the liver diverticulum. Subsequently, the epithelial cells of the liver diverticulum proliferate, forming cords which invade the loose mesenchyme of the ventral septum transversum (Le Douarin 1975; Zaret 1996, 1998). As development proceeds the foetal liver becomes increasingly structured. The endodermal chords anastomose with the endodethelial capillary bed connected to the nearby vitteline veins. These mesenchymally derived tissues will go on to line the hepatic blood sinusoids (Severn, 1972). By 12 d.p.c., hepatocytes derived from the endodermal cells of the liver diverticulum contribute around 40% of the total cells of the foetal liver (Jones, 1970).

From mid to late gestation, the liver becomes the major site of haematopoiesis in the mouse embryo. Cells of the erythroid lineage first colonise the foetal liver around 10 d.p.c. and by 11.5 d.p.c. myeloid cells, macrophages, B cells and presumptive

haematopoietic stem cells are also present (Medvinsky *et al.*, 1993; Dzierzak and Medvinsky, 1995). The haematopoietic cells are mesodermal in origin deriving from progenitor cells migrating from both the yolk sac mesoderm and the aorta-gonad-mesonephrous (AGM) region (Medvinsky *et al.*, 1993). As development progresses, haematopoietic cells are more commonly found extravascularly in the foetal liver with numbers progressively decreasing as haematopoiesis switches to the bone marrow of the neonate.

1.4.2. Tissue Interactions Mediating Liver Development

Tissue explantation and transplantation studies, classically in the chick and more recently in the mouse, have uncovered a series of interactions between endodermally derived epithelia and ventral mesenchyme derived from mesoderm necessary for liver ontogeny (Zaret, 1996, 1998). Such mesenchymal-epithelial interactions driving organogenesis is a common developmental mechanism observed in a number of other organs including kidney, lung, and mammary gland (Birchmeier and Birchmeier, 1993).

In the chick, endoderm from the 5-somite stage onwards has the ability to differentiate into phenotypically distinct hepatocytes upon transplantation into host embryos. Prior to this stage, hepatogenesis is only observed if the endoderm is explanted with precardiac mesoderm suggesting that initial hepatic determination of endoderm is dependent on its interaction with precardiac mesoderm (Le Douarin, 1975). Differentiation of determined hepatic endoderm from the 5 somite stage onwards *in vivo* or *in vitro* is dependent on the hepatic mesenchyme. Interestingly, only hepatic endoderm is responsive to inductive mesenchyme but other mesenchymes derived from lateral plate mesoderm are capable of inducing hepatic endoderm differentiation (Le Douarin, 1975).

Presumptive hepatic endoderm isolated from the ventral foregut from different embryonic stages of the mouse was cultured with either chick hepatic or pulmonary

mesenchyme and then grafted into chick (Houssaint, 1980). From the 9 somite onwards, the murine foregut endoderm had the ability to differentiate into morphologically distinct hepatocytes expressing the hepatic marker α-foetoprotein (AFP). Moreover, these studies showed that both hepatic and pulmonary lateral plate mesenchyme could support the differentiation and proliferation of endodermally derived hepatocytes correlating well with the requirements of chick hepatic endoderm. In the absence of co-cultured mesenchyme mouse endoderm failed to differentiate (Houssaint, 1980; Le Douarin, 1975).

Gualdi *et al.*, (1996) showed that ventral foregut endoderm explants from 4-6 somite embryos would only express the hepatocyte specific markers AFP and serum albumin (*alb*) when cultured alongside cardiac mesoderm consistent with previous studies (Houssaint, 1980; Le Douarin, 1975). Perhaps more interesting was the finding that more posterior dorsal endoderm was capable of expressing AFP and *alb* when not in contact with adjacent dorsal mesoderm and ectoderm. Moreover, this posterior dorsal mesoderm and ectoderm was capable of inhibiting the expression of AFP and *alb* in explants of ventral endoderm and cardiac mesoderm (Gualdi *et al.*, 1996). This indicates that gut endoderm has the potential to differentiate along the hepatic lineage, an ability which is normally restricted by adjacent ectoderm and mesoderm.

In a recent study by Jung *et al.* (1999), tissue explant studies similar to those performed by Gualdi *et al.*, (1996) were used to determine the molecules mediating the inductive signals responsible for the specification of hepatic endoderm by the cardiac mesoderm. The treatment of isolated foregut endoderm with either fibroblast growth factor (FGF) 1 or FGF2 was sufficient to induce the liver specific gene expression effectively replacing the effect of cardiac mesoderm in comparable cultures.

These studies allow us to postulate a programme of interactions necessary for the determination and early differentiation of the foetal liver *in vivo* (Figure 1.6). Hepatic determination occurs in the 4-8 somite embryo (8-8.5 d.p.c.) via interactions between the cardiac mesoderm and the epithelium of the ventral foregut endoderm

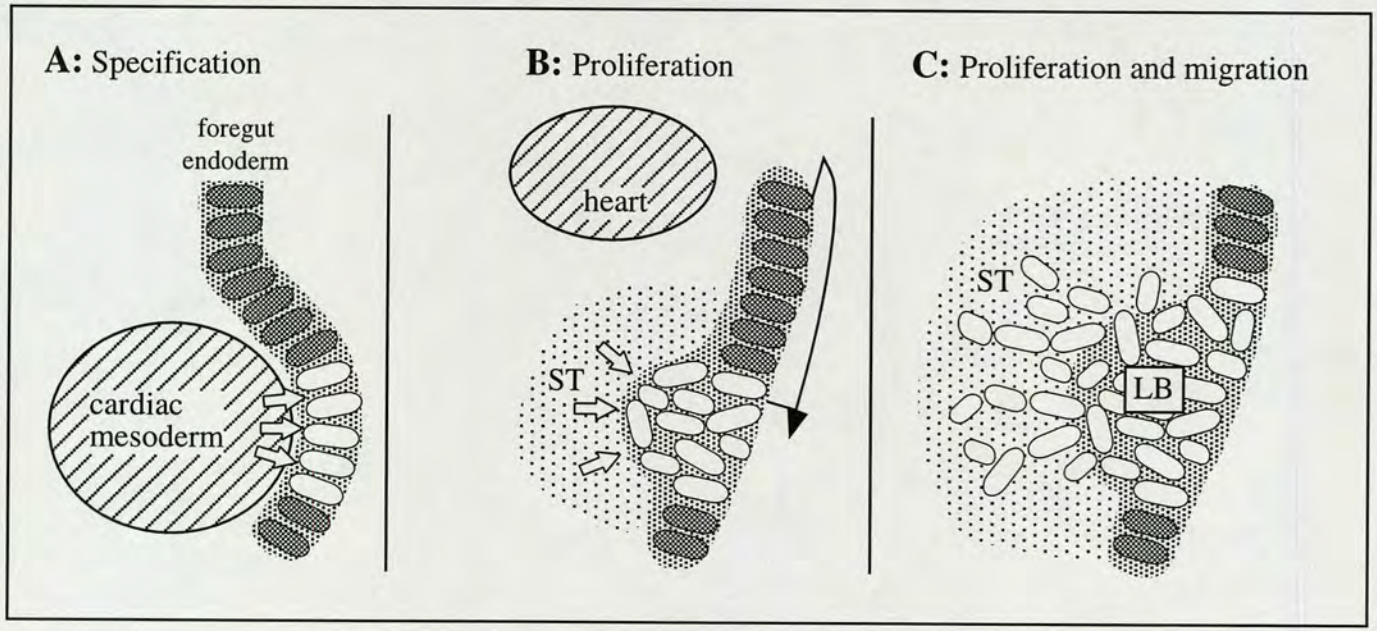


Figure 1.6: Inductive interactions during the formation of the foetal liver

A: Between the 4 and 8 somite stage of embryogenesis, inductive signals from the cardiac mesoderm activate liver specific genes in the foregut endoderm specifying this region to differentiate along the hepatic lineage (shown as white blocks).

B: From the 9 somite stage foregut closure proceeds posteriorly drawing the hepatic endoderm away from the developing heart (black arrow). During this process signals (white arrows) from the septum transversum (ST) induces the hepatic endoderm to proliferate within the endodermal layer.

C: From 9.5d.p.c. the proliferating early hepatocytes migrate into the loose mesenchyme of the septum transversum where they anastomose with capillaries forming the liver bud (LB).

(Figure 1.6A). After specification, foregut closure and extension pulls the hepatic endoderm more posteriorly as interactions with the adjacent mesenchyme of the septum transversum induces proliferation of the hepatic endoderm (Figure 1.6B). At 9.5 d.p.c. continued mesenchymal-epithelial interactions drives the migration of endodermal chords into the septum transversum forming the liver bud (Figure 1.6C).

1.4.3. Hepatic Markers

It is of interest to highlight at this stage the two related serum proteins AFP and *alb* which have been used as definitive markers of hepatocyte differentiation in recent studies (Gualdi *et al.*, 1996; Houssaint, 1980; Cascio and Zaret, 1991; Shiojiri, 1981). From around 7 d.p.c., AFP is expressed at high levels in the endodermal cells of the visceral yolksac (Dziadek and Adamson, 1978; Dziadek and Andrews, 1983) with *alb* expression first detected at 13.5 d.p.c. in the same tissue (Meehan *et al.*, 1984). Low levels of AFP and *alb* transcripts can be first detected in the presumptive hepatic endoderm between 8.0-8.5 d.p.c. prior to formation of the liver diverticulum (Shiojiri, 1981; Gualdi *et al.*, 1996). From 9.5 d.p.c. onwards, high levels of AFP are observed in the liver (Shiojiri, 1981; Dziadek and Andrews, 1983) with additional expression in the developing gut (Tyner *et al.*, 1990). AFP expression shuts down at birth with re-expression in the adult associated with tumour formation (Abelev *et al.*, 1971). Similarly a higher level of *alb* expression is seen at 9.5 d.p.c. in the liver diverticulum and by 12 d.p.c. around a 20 fold increase in expression is observed which is maintained throughout gestation and in the adult (Cascio and Zaret., 1991). Despite their extensive use as markers for both visceral endoderm and hepatic endoderm, the exact function of *alb* and AFP is unknown although many functions have been postulated (Uriel *et al.*, 1976; Brenner *et al.*, 1980; Leffert *et al.*, 1978).

1.4.4. Molecular Basis of Hepatic Determination

With such a significant amount of work carried out defining the tissue interactions necessary for the transition of the foregut endoderm to the hepatic lineage, it is of great interest to characterise the molecular elements mediating these interactions. Although the expression of a number of factors in the definitive endoderm and liver primordia suggested roles in determination, gene knock-out analysis has failed to define functions for these factors in hepatic specification. Nevertheless, more recent studies are implicating two of these factors, HNF-3β and GATA-4, in potentiating hepatic determination.

The transcription factors HNF-3 and HNF-4 were initially isolated as factors binding to the upstream regulatory elements of genes expressed in adult hepatocytes (Costa *et al.*, 1989).

HNF-4 is a member of the steroid hormone receptor superfamily (Sladek *et al.*, 1990). Expression of HNF-4 is first detected in the visceral endoderm of the yolk sac at 4.5 d.p.c. and is the sole site of expression until 8.5 d.p.c. when expression is seen in the liver diverticulum. From this stage, expression is seen at high levels in the proliferating hepatocytes with additional expression in the gut and nephrogenic tissue (Duncan *et al.*, 1994). Mutational analysis of HNF-4 results in embryonic lethality around 7.5 d.p.c. with the mutant embryos displaying increased cell death in the embryonic ectoderm and impaired gastrulation (Chen *et al.*, 1994a). Consequently, such an early phenotype, prior to liver determination, makes it impossible to assess the potential involvement of HNF-4 in liver ontogeny.

HNF-3 is the defining member of the winged helix transcription factor gene family conserved throughout evolution which includes the *forkhead* gene in *Drosophila* (Lai *et al.*, 1993). In the mouse embryo, expression of HNF-3β is first observed at gastrulation in the anterior portion of the primitive streak. At the headfold stage, HNF-3β expression can be seen in the node, notochord as well as in the floorplate and definitive endoderm where its expression overlaps with the related gene HNF-3α (Ang

et al., 1993; Sasaki *et al.*, 1993; Monaghan *et al.*, 1993). At 8.5 d.p.c. HNF-3α and HNF-3β are expressed throughout the gut endoderm with high levels observed in the presumptive hepatic endoderm at 8.5 d.p.c. and in the liver diverticulum from 9.0 d.p.c. (Ang *et al.*, 1993). Disruption of HNF-3β causes embryonic lethality with defects in node and notochord formation which subsequently affects dorsal-ventral patterning. Furthermore, HNF-3β mutants form definitive endoderm but gut morphogenesis is disrupted with no invagination of the foregut pocket (Ang *et al.*, 1994).

Defects in foregut morphogenesis are also observed in embryos deficient in the zinc finger transcription factor GATA-4 (Kuo *et al.*, 1997; Molkentin *et al.*,1997). Expression of GATA-4 is seen in the visceral endoderm, developing heart and gut endoderm (Arceci *et al.*, 1993; Molkentin *et al.*, 1997). Inactivation of GATA-4 affects rostral-to caudal and lateral-to-ventral morphogenesis with the ventral closure of the yolk sac and the formation of the heart tube and foregut pocket disrupted (Kuo *et al.*, 1997; Molkentin *et al.*,1997). Following on from this study, Narite *et al.* (1997a, 1997b) has shown that the ventral pattering defect is a result of GATA4 function, either in the visceral endoderm and/or definitive endoderm and is not intrinsic to cardiac tissue.

Both HNF-3β and GATA-4, therefore, play a role in foregut morphogenesis, a function which is conserved between structural homologues in other metazoans. In *Drosophila*, the HNF-3β related gene *forkhead* and the GATA-like gene *serpent* are required for endodermal gut development (Weigel *et al.*,1989; Rehorn *et al.*, 1996). *C.elegans* also possess a GATA factor *end-1* which is essential for endoderm formation (Zhu *et al.*, 1997).

Although both HNF-3β and GATA-4 mutants present phenotypes before hepatic specification, *in vivo* footprinting studies of the serum albumin enhancer element has suggested roles for both proteins in hepatic potentiation. HNF-3β and GATA-4 bind distinct sites in the *alb* enhancer element (which drives liver specific *alb* expression) in undifferentiated gut endoderm with the potential to activate albumin

expression. Subsequently, the binding of other factors to additional enhancer sites is observed upon hepatic specification (Gualdi *et al.*, 1996; Bossard and Zaret, 1998). From these observations it is postulated that these two factors act to potentiate the expression of hepatic specific genes (Zaret, 1998; Gualdi *et al.*, 1996; Bossard and Zaret, 1998).

Another candidate gene which may be involved in hepatic specification is the divergent homeobox gene *Hex*. Expression of this gene is observed in the visceral endoderm and then later in the earliest definitive endoderm and subsequently at high levels in the ventral foregut and the liver bud (Thomas *et al.*,1998a).

1.4.5. Mutational Analysis of Liver Development

The molecules and pathways underpinning the interactions between the hepatic endoderm and ventral mesenchyme after its specification are being illuminated by a number of gene mutation studies which are outlined below.

The secreted glycoprotein scatter factor/ hepatocyte growth factor SF/HGF was predicted to be involved in mesenchymal-epithelial interactions by its defining effects on motility and morphology of a variety of epithelial cells in culture (Stoker *et al.*, 1987; Gherardi *et al.*, 1989, Montesano *et al.*, 1991) and its expression in the mesenchyme of various developing organs (Sonnenberg *et al.*, 1993). In addition, SF/HGF was shown to be a powerful mitogen of primary hepatocytes in culture (Nakamura *et al.*, 1989) and may have a role *in vivo* on the proliferation of hepatocytes in response to liver injury (Zarnegar *et al.*, 1989). Targeted disruption of SF/HGF results in embryonic lethality from 14-16 d.p.c. Affected embryos have hypoplastic livers with disrupted architecture containing a reduced number of morphologically abnormal hepatocytes. The mesenchymally derived sinusoidal cells of the livers which express SF/HGF are unaffected (Schmidt *et al.*, 1995).

Complementing this study, animals lacking the receptor for SF/HGF, the receptor tyrosine kinase c-*met*, which is expressed at high levels in the hepatic

endoderm, display a strikingly similar phenotype to the SF/HGF mutant embryos (Bladt *et al.*, 1995).

The divergent homeobox gene *Hlx* was identified from its expression in restricted haematopoietic lineages (Allen *et al.*, 1991). During development *Hlx* expression is restricted to mesodermally derived tissues including the mesenchyme of the liver, gall bladder and intestines during organogenesis (Lints *et al.*, 1996). Targeted disruption of *Hlx* results in a perturbation in the outgrowth of the endodermally derived liver diverticulum as well as in the elongation and looping of the intestines (Hentsch *et al.*, 1996).

The proto-oncogene *c-jun* is a component of the transcription factor AP-1, an effector of the protein kinase C signalling pathway which is widely expressed throughout the embryo and adult (Ransone and Verma, 1990; Angel *et al.*, 1987; Lee *et al.*, 1987). Mice deficient in *c-jun* die at mid to late gestation with a severely hypoplastic liver and impaired hepatic erythropoiesis (Hilberg *et al.*, 1993).

The four studies outlined above highlight similar phenotypes relating to the development of the liver, that is, defects intrinsic to the outgrowth of the hepatic endoderm. Significantly, both SF/HGF and *Hlx*, are expressed in the mesenchyme of the septum transversum and represent factors which could mediate the dependency of the hepatic endoderm on adjacent mesenchyme for organogenesis. Moreover, the studies also implicate certain signalling cascades involved in hepatogenesis. SF/HGF has been shown to activate AP-1 (*c-jun*) via the JNK1 signalling pathway in cultured hepatocytes and fibroblasts (Auer *et al.*, 1998; Rodrigues *et al.*, 1997). Although this is not the sole pathway either activated by SF/HGF or activating c-*jun*, it does provide a link to the comparable phenotypes observed when these different molecules are mutated. By this rationale, it is not unreasonable to suggest that the transcription factor *Hlx* may function to activate SF/HGF in the hepatic mesenchyme although there is no direct evidence for this (Hentsch *et al.*, 199).

NF-kB is a transcription factor which has been predicted to be involved in a wide range of cellular functions but shows a liver specific defect when mutated (not

unlike *c-jun*). This factor consists of the polypeptides RelA/p65 and p50 and is believed to regulate infection, inflammation and stress (Baeuerle and Henkel, 1994). Mice null for the RelA subunit display a massive degeneration of the foetal liver mediated by the apoptosis of hepatocytes resulting in lethality around 15 d.p.c. (Beg *et al.*, 1995).

Similar phenotypes are observed in mice deficient in *jumonji* and N-*myc*. *Jumonji*, initially identified by gene trapping, is expressed in the stromal cells of the liver with mutant embryos displaying hypoplastic livers, thymus and spleen with increased hepatocyte apoptosis in the liver by 13.5 d.p.c. (Motoyama *et al.*, 1997). N-*myc* mutants show growth retardation in many tissues, most notably the heart, as well as increased apoptosis of hepatocytes by 11.5 d.p.c. (Giroux *et al.*, 1998).

Both *jumonji* and N-*myc* are not expresssed in hepatocytes and yet their functions are absolutely required for hepatocyte survival. The mutant phenotypes therefore are postulated to be a manifestation of the dependency the liver has to continued cellular interactions throughout its ontogeny (Zaret, 1998).

1.4.6. Summary

The pioneering studies of Le Douarin over 20 years ago are only now beginning to be compounded by gene knock-out studies identifying factors and pathways which may be involved in the tissue interactions mediating liver ontogeny. Most success has been in identifying factors involved in the proliferation of the liver after specification. Notably, however, there has been a failure to identify molecules specifying the hepatic endoderm. Consequently, it is important to identify molecules expressed early and exclusively in the hepatic lineage as potential affectors or effectors of specification.

1.5. Experimental Approach

The aim of my project is to characterise the I114 gene trap cell line. I114 was initially selected after showing an induction in reporter activity using the RA pre-screening strategy of Forrester *et al.*, (1996) outlined in Section 1.3.2.5(iv). In I114 chimaeric embryos reporter expression was identified in the foetal liver of between 9.5-10.5d.p.c. (Forrester *et al.*, 1996). The expression pattern of the reporter gene prompted the characterisation of the I114 line as a means of identifying a gene with a potential role in liver development. Chapter 3 describes the exquisite reporter expression profile of the I114 gene trap integration which exclusively marks the specification and proliferation of the foetal liver. Chapter 4 identifies a complex integration event associated with this reporter activity and Chapter 5 and 6 resolve the integration event to identify the endogenous gene responsible for the liver specific expression pattern.

Chapter 2

MATERIALS AND METHODS

Unless otherwise stated, analytical grade chemicals were obtained from either Sigma or BDH Laboratory supplies (Merc Ltd) with the exception of absolute ethanol (Hayman Ltd, Litham, UK). Analytical grade agarose was supplied by GibcoBRL with NuSieve low melting point (LMP) agarose supplied by Flowgen (Sittingbourne, UK). All bacterial media components were supplied by DIFCO laboratories. Synthetic oligonucleotides were synthesised by OSWELL DNA Service (University of Southampton, UK). Radioisotopes were supplied by Amersham International plc (Little Chalfont, UK) or NEN and X-ray film (Kodak XOMAT XAR-5) supplied by IBI Ltd (Cambridge, UK).

2.1. Molecular Biology Methods

General molecular biology techniques and the preparation of standard solutions was carried out according to Sambrook *et al.* (1989) unless otherwise stated. Restriction enzyme digests were performed as recommended by the suppliers (Boehringer Mannheim). Digest products were routinely analysed by agarose gel electrophoresis with gels cast and ran in 1xTAE buffer at appropriate concentrations containing 0.5µg/ml ethidium bromide. 0.5µg of 1kb DNA ladder (GibcoBRL) was loaded on each gel as a size standard.

2.1.1. General Cloning Techniques

DNA fragments and RACE-PCR products (Protocol 2, Section 2.1.8.1.) were routinely subcloned by restriction digestion and cloning into the plasmid pBluescript

KS (Stratagene). The Zero Background™ cloning kit (Invitrogen) was also used to clone certain restriction fragments. Pfu polymerase generated blunt end RACE-PCR products were subcloned using the Zero Blunt™ PCR cloning kit (Invitrogen) and Taq polymerase PCR products (which have a deoxyadenosine added to the end) were subcloned using the TOPO TA Cloning® kit (Invitrogen). All the kits were used as instructed by the manufacturers.

2.1.1.1. Gel Purification of DNA

Restriction fragments for subcloning and the pBluescript vector were digested and purified as follows:

Restriction digest products were run on 1.0% low melt agarose gels until suitable separation had occurred to isolate a single band. Bands of interest were excised from the gel using a scalpel under long wavelength UV illumination. The gel slice was weighed and an equal volume of dH_20 added. Sodium chloride was added to a final concentration of 0.1M. The sample was heated to 70°C for 5-10 minutes in a hot block (Techne) and vortexed to melt the agarose. An equal volume of Tris-saturated phenol (Fisons) prewarmed to 37°C was added to the gel solution and the sample vortexed for 1-2 minutes. The suspension was then centrifuged at 13,000 x g. for 3 minutes and the upper aqueous phase removed and extracted in an equal volume of phenol/chloroform/IAA (25:24:1) which was vortexed and centrifuged at 13,000 x g. for 1 minute. The aqueous phase was removed, extracted with an equal volume of chloroform and centrifuged at 13,000 x g. for 30 seconds. The aqueous phase was precipitated with 2 volumes of 100% ethanol at -20°C for a minimum of 10 minutes and DNA pelleted by centrifugation at 13,000 x g. for 10 minutes. The pellet was washed with 70% ethanol, air dried and resuspended in 20µl of T.E. (10mM Tris-HCl pH8.0; 1mM EDTA pH8.0)

2.1.1.2. Ligations

For ligations, pBluescript was digested and gel purified (Section 2.1.1.1). Restriction fragments for subcloning into pBluescript were either gel purified or, after digestion, extracted with an equal volume of phenol/chloroform, precipitated with 0.1 volume 3M NaOAc (pH5.2) / 2 volumes of 100% ethanol at -20°C for a minimum of 10 minutes and resuspended in appropriate volume of sterile water.

Ligation reactions were set up with approximately a 1:3 vector to insert molarity ratio with 1µl (1unit) of T4 DNA Ligase (Boehringer) and 1µl 10x ligation buffer (660mM Tris.HCl pH7.5; 50mM $MgCl_2$; 10mM DTT; 10mM ATP; Boehringer) to a total volume of 10µl. Cohesive-end ligations were performed at 16°C for a minimum of 1 hour, blunt-end ligations were incubated overnight at 16°C. To determine the background level of vector re-ligation, a control reaction containing vector alone was also set up. Ligation reactions were precipitated with 0.1 volume of 3M NaOAc (pH5.2)/ 2 volumes of 100% ethanol, washed in 70% ethanol and resuspended in 4µl of T.E prior to transformation.

2.1.1.3. Transformation of Bacterial Cells

Electroporation was routinely used to transform the bacterial strain DH5α which were prepared as follows.

A single bacterial colony from a freshly streaked plate was used to inoculate 25 mls Luria broth (LB) culture media (1% w/v Tryptone (Difco); 0.5% w/v yeast extract (Difco); 85mM NaCl) which was grown at 37°C overnight with shaking. 5mls of this overnight culture was used to inoculate 50mls LB broth and the culture grown to OD_{600} = 0.3 to ensure the cells are in log phase. The cells were then chilled on ice for 15 minutes, centrifuged at 4000 x g. for 10 minutes at 4°C. In all subsequent steps, centrifuge bottles, tubes, pipettes and strerile water are all kept at 4°C. The bacterial pellet was resuspended in 500mls of ice cold water, re-centrifuged as before and resusupended in 250ml of ice cold water. Centrifugation was repeated and the bacterial pellet resuspended in 40ml of ice cold 10% glycerol. After a final centrifugation the

pellet was resusupended in 1ml of ice cold 10% glycerol. The electrocompetent bacterial cells were dispensed into 100µl aliquots in prechilled eppendorf tubes and snap frozen in liquid nitrogen before storage at -80°C.

Unless using fresh, aliquots of electrocompetent cells were thawed on ice for ~30 minutes. 40 µl of electrocompetent cells were mixed with 4µl of the precipitated and resuspended ligation reaction in an eppendorf tube on ice. The sample was transferred to a prechilled cuvette with a 1mm electrode gap (Biorad) and the cells were electroporated (25 mFD, 200 ohm and 1.8 kV using Biorad's Gene Pulser) which would typically produce a time constant of 4.5. 1.0ml of LB broth at room temperature was immediately added to the cells and the culture transferred to an eppendorf tube which was incubated at 37°C for 45-60 minutes to allow the cells to recover. Appropriate dilutions were plated onto LB agar plates (1.5% w/v agar in LB) with 100µg/ml Ampicillin as selection. When blue/white selection was also required, standard 9cm agar plates were treated with 40µl of 25mg/ml X-gal (5-bromo-4-chloro-3-indolyl-β-D-galactopyranoside) in dimethyl formamide (DMF) and 4µl of 0.2mg/ml IPTG (isopropylthio-β-D-galactoside) prior to plating.

2.1.1.4. Screening Transformants

(a) Colony Lifts

Analysis of a large numbers of recombinant clones on agar plates prior to plasmid preparation was carried out by taking colony lifts onto nylon membranes, lysing the colonies and hybridising to a radiolabelled probe.

A 9cm diameter Hybond N+ nylon membrane (Amersham) was placed on top of the agar plate and asymmetric orientation marks made using a needle. The membrane was removed from the plate and placed, colony side up into a 1ml puddle of denaturing solution (1.5M NaCl; 0.5M NaOH) on SaranWrap for 2 minutes. The membrane was removed and placed colony side up in a 1ml puddle of neutralising solution (0.5M Tris.HCl pH8.0; 1.5M NaCl) for 5 minutes. The membrane was then rinsed

(submerged) in 2xSSC. The membrane was then air dried briefly and baked at 120°C for 30 minutes after which it hybridised to the appropriate probe.

(b) Colony Screening by PCR

Large numbers of recombinant bacterial clones were screened for inserts by PCR primers flanking the insertion site of the cloning plasmid (T3 and T7 for pBluescript and SP6 and T7 for pCR®-Blunt and pZErO™-2).

Bacterial colonies were picked using a yellow pipette tip into 20µl of water. 7µl of the culture was stored at 4°C and the remaining 13µl of the culture was added to 2µl 10x PCR Buffer buffer (50mM KCl/10mM Tris (pH8.3)); 2µl 25mM $MgCl_2$; 0.4µl 10mM dNTPs; 1µl 100ng primer x; 1µl 100ng primer y; 0.2µl (1 unit) Taq DNA polymerase (Promega). The samples were overlaid with mineral oil and incubated at 94°C for 5 minutes and taken through 25 PCR cycles of 94°C for 30 secs; 55°C for 30 secs; 72°C for 45 secs. Half the PCR reaction was separated on an agarose gel, Southern blotted and hybridised to the relevant probe. Positive colonies were grown for plasmid preparation using the 7µl of stock culture.

2.1.2. Isolation of Nucleic Acids

2.1.2.1. Plasmid Preparation

Plasmid isolation was either performed using the Qiagen plasmid isolation kit according to the manufacturers instructions or using the following protocol. The amounts used are for minipreps of plasmids. The equivalent amounts used for midipreps and maxipreps are provided in parentheses.

A single colony was picked with a sterile yellow Gilson tip into 2mls (MID- 25mls; MAX- 100mls) of LB with the appropriate antibiotic and the culture incubated at 37°C overnight with shaking. In an eppendorf tube, 1.5mls of overnight culture centrifuged for 3 minutes at 13,000 x g. to pellet the cells (MID and MAX- 6,000 x g for 15 mins). The supernatant was discarded and the pellet resuspended in 200µl (MID- 4mls; MAX- 10mls) of resuspension buffer (50 mM Tris-HCl (pH 8.0); 10 mM

EDTA; 100µg/ml RNase A- stored at 4°C). 200µl (MID- 4mls; MAX- 10mls) of the freshly prepared lysis buffer (0.2 M NaOH;1% SDS) was added and the tube inverted gently several times and the left on ice for 5 minutes. 200µl (MID- 4mls; MAX- 10mls) of ice cold neutralisation buffer (3M KAcetate pH4.8 with acetic acid) was added the tube inverted gently several times and then left on ice for 5 minutes. The samples were then centrifuged at 13,000 x g for 5 minutes to pellet the precipitate (MID and MAX- 13,000 x g in Sorval SS-34 rotor for 15 minutes using Oakridge tubes). The supernatant (6-700µl for miniprep) was removed and added to 0.7 volumes of isopropanol in a fresh eppendorf tube (MID and MAX- supernatant was mixed with isopropanol in Corex glass tubes. If the supernatant is not clear, it can be passed through a 0.45µm syringe filter- Nalgene). The DNA is pelleted by centrifugation at 13,000 x g for 10 minutes, washed in 70% ethanol, air dried and resuspended in 50µl of T.E. (MID and MAX- centrifugation at 11,000 x g in Sorval SS-34 rotor and resuspended in 500µl)

DNA from this preparation was of sufficient quality for diagnostic digest analysis. If the miniprep DNA preparation was to be used in the automatic sequencing protocol (Section 2.1.7.2.) and for all midi and maxipreps, the DNA was purified further by adding an equal volume of 13% PEG-8000 (in 1.6M NaCl) to the DNA and incubated on ice for 1 hour. The DNA was pelleted by centrifugation at 13,000 x g for 5 minutes, washed twice with 70% ethanol, air dried and resuspended in an appropriate amount of either TE or sterile water (if used for automatic sequencing).

2.1.2.2. PAC Preparation

PAC clones were provided by the UK HGMP Resource Centre (http://www.hgmp.mrc. ac.uk) after the screening of 7 gridded filters of the mouse genomic PAC library RPC121 with the L-69 (Group II) fusion transcript. The PAC clones were streaked out to single colonies on LB with 25µg/ml kanamycin. PAC clones were prepared using the Qiagen Midiprep and Maxiprep Kits according to the manufacturers instructions with the following modifications:

For the midiprep protocol, 100mls of overnight culture (LB + 25µg/ml kanamycin) was used. 8mls (2x recommended amount) of each of the buffers P1, P2, P3 were added (steps 1-3). For the elution of the plasmid from the Qiagen column (step 9), the elution buffer (QF) was heated to 65°C and added to the column 1ml at a time.

For the maxiprep protocol, 250mls of overnight culture (LB+ Kanamycin) was used. 20mls (2x recommended amount) of each of the buffers P1, P2, P3 were added (step 1-3). For the elution of the plasmid from the Qiagen column (step 9), the elution buffer (QF) was heated to 65°C and added 2mls at a time.

Restriction enzyme digestion of PAC clones was carried out as for plasmids.

2.1.2.3. Isolation of Genomic DNA

(a) ES Cells:

ES cells grown to confluency in $25cm^2$ flasks were rinsed twice with PBS (137mM NaCl, 2.7mM KCl, 4.3mM Na_2HPO_4 and 1.4mMKH_2PO_4) and lysed overnight at 55°C in 5ml of lysis buffer (100mM Tris.HCl pH8.5; 5mM EDTA; 0.2% SDS; 200mM NaCl; 100µg/ml Proteinase K (Sigma)-added fresh). The lysate solution was extracted with equal volumes of phenol/chloroform/isoamyl alcohol (25:24:1) and centrifuged at 6,000 x g or 10 minutes. The aqueous phase was decanted and the extraction repeated with chloroform/isoamyl alcohol (24:1). The aqueous phase was precipitated using an equal volume of isopropanol. The DNA was spooled out of the tube using a sterile gilson tip, rinsed in 70% ethanol, and resuspended in 50µl of T.E. overnight at 4oC and the DNA concentration determined by measuring the OD at 260nm.

(b) Tissue:

Tail tips (0.5-1.0cm) from newly weaned animals was digested overnight at 55°C in 0.5ml lysis buffer (100mM Tris.HCl pH8.5; 5mM EDTA; 0.2% SDS; 200mM NaCl; 100µg/ml Proteinase K added fresh) with agitation. The lysate was vortexed and

centrifuged at 13,000 x g for 10 minutes to remove bones and hair. The resulting supernatant was extracted and precipitated as described above for ES cells.

2.1.2.4. Isolation of High Molecular Weight Genomic DNA

ES cells were grown to confluency in a 175cm^2 flask, medium removed and cells rinsed in PBS. Cells were trypsinised to a single cell suspension, centrifuged for 5 minutes at 1,200 x g. and resuspended in PBS. The cells were counted in a haemocytometer, spun amd resuspended in PBS to a final concentration of 2.5x10^7 cells/ml. 2 volumes of cells were mixed with 3 volumes of 1% LMP agarose in PBS (kept at 37^{o}C). Immediately, 100µl of the cell suspension was dispensed into a perspex plug mold (Biorad) with approximately 6 x 2 x 10mm size wells on ice. Once the agarose plugs had set, the plugs were pushed out into 50ml tube (Corning) containing NDS (0.5M EDTA; 1% sodium lauroyl sarcosine; 0.01M Tris pH 7.5) with 1mg/ml proteinase K and were incubated overnight at 50^{o}C. The plugs were then rinsed twice in NDS for 2 hours and stored in NDS at 4^{o}C. The plugs are stable for at least 1 year at 4^{o}C.

2.1.2.5. Isolation of RNA

Total RNA was isolated using a protocol modified from Chomczynski and Sacchi, (1987). ES cells for RNA preparation were grown to confluency in either 25cm^2 flasks or 9cm plates. Cells were rinsed twice with PBS and then harvested by pipetting up and down with 5mls of Solution D (4.4M guanidinium isothiocynate; 25mM sodium citrate; 0.6% sarcosyl; 100mM β-mercaptoethanol). The lysate was transferred to a Falcon 2059 centrifuge tube. For RNA isolation from tissues, approximately 100-200mg of freshly dissected tissue was homogenised using a standard ground glass homogeniser in 5mls of Solution D. The lysate was transferred to a Falcon 2059 tube. The following protocol was used to isolate RNA from ES cells and tissue.

To 5mls of the Solution D lysate, 0.5mls 2M Sodium Actetate (pH4.0), 5mls Phenol (T.E. or Tris-saturated-Fisons) and 2mls chloroform were sequentially added. The samples were put on ice for 15 minutes and then centrifuged at 6,000 x g for 15 minutes at 4°C (Sorval SS34 rotor). The upper aqueous phase containing the RNA was removed and precipitated in an equal volume of isopropanol (~6mls) at 4°C for 1 hour. The RNA was pelleted by centrifugation at 8,000 x g for 20 minutes at 4°C (Sorval SS34 rotor). The pellet was resuspended in 0.3ml Solution D, reprecipitated with an equal volume of isopropanol at 4°C for 1 hour. Late gestation or adult liver samples with a high glycogen content were precipitated at 4°C overnight with an equal volume of 4M lithium chloride to prevent co-precipitation of glycogen (this was only if the RNA was not being used in reverse transcription as Cl^- ions inhibit RT). The precipitates were pelleted by centrifugation at 13,000 x g for 10 minutes at 4°C and the pellet washed with 70% ethanol, supernatant removed with a P2 Gilson tip and the pellet immediately resuspended in 50µl of DEPC treated water and left at 4°C overnight. The RNA concentration was determined by absorbance at 260nm. The ratio of 260nm/280nm was used to estimate the purity of the sample with a ratio close to 1.8 indicative of low protein contamination. The samples were stored long term at -70°C.

2.1.3. Analysis of High Molecular Weight DNA

2.1.3.1. Digestion of Genomic DNA Agarose Plugs

High molecular weight agarose plugs (Section 2.1.2.4.) were cut into three sections each sufficient for an single restriction digest. Agarose plugs were rinsed twice in TE at 4°C for a minimum of 30 minutes per wash (the first wash was normally performed overnight). The plugs were then washed at 4°C for 30 minutes in TE containing 0.1M phenylmethylsulfonyl fluoride (PMSF) and subsequently rinsed twice in TE for 30 minutes. The plugs were then rinsed in the appropriate 1x concentration of restriction buffer for 30 minutes. Each plug was digested overnight at 37°C in a volume of 70 µl containing 1x restriction buffer, spermidine (2mM spermidine for

restriction buffers with 50-100mM salt and 5mM spermidine for buffers with over 100mM salt), 200µg/ml BSA and 2µl of 40u/µl restriction enzyme. A further 1µl of restriction enzyme was added and the plugs digested for 2 hours to overnight. The restriction digests were stopped by adding 100µl of 0.5M EDTA and the samples were ready for analysis by pulse field gel electrophoresis (Section 2.1.3.2.). To control for degradation of genomic DNA, a control plug is taken through the same protocol with the omission of the restriction enzyme.

2.1.3.2. Pulse Field Gel Electrophoresis (PFGE)

Separation of high molecular weight genomic DNA fragments was carried out using Biorad's CHEF-DR II pulse field nucleic acid electrophoresis system.

A 0.8% agarose gel in 0.5xTBE was poured using the CHEF-DR II casting stand and wells. Digested genomic DNA and molecular weight marker agarose plugs were inserted into the wells of the gel and sealed with 1%LMP agarose. Liquid samples (PAC restriction digests) were mixed in an equal volume of 1% LMP agarose and loaded into the wells. (Molecular Weight Markers- Mid range I PFG marker and λ ladder; New England Biolabs). Once the wells had set the gel was submerged in 2 liters of 0.5% TBE in the electrophoresis cell at 14^0C. The gel was then subjected to a switch time ramp of 1-20 second pulses for 20-24 hours at a constant 6V/cm. The switch time ramp increases the mobility of the DNA fragments as a function of molecular weight by gradually changing the switch times through the course of a run. This provides a more linear separation of DNA fragments through a range of molecular sizes. After electrophoresis, the pulse field gel was stained with 10µl of ethidium bromide (25mg/ml) in 500ml of 0.5x TBE for 30 minutes. The gel was rinsed twice in sterile water, photographed and Southern blotted (Section 2.1.4.1.)

2.1.4. Nucleic Acid Transfer to Membranes

2.1.4.1. Southern Blotting

Southern blotting as described in Sambrook *et al.*, 1989 was carried out for the transfer of high molecular weight DNA from the PFGE of restriction digested genomic DNA and PAC DNA. Prior to blotting, PFGE gels were stained in ethidium bromide (~1μg/ml), rinsed in water and photographed.

High molecular weight DNA was depurinated (acid nicked) in 0.25M HCl for 20 minutes with gentle agitation. The gel was rinsed in water and then denatured in 0.5M NaOH, 1.5M NaCl for 30 minutes and neutralised for 30 minutes in 1.5M NaCl, 1mM EDTA, 0.5M Tris-HCl (pH7.2). A toughened glass plate was placed across a Pyrex dish containing 20x SSC (3M NaCl, 0.3M trisodium citrate) and two strips of Whatman 3MM paper were placed on top of the glass plate with the ends submerged in 20x SSC to act as a wick. The gel was placed on top of the paper and surrounded by parafilm to prevent evaporation of the buffer. Hybond N+ nylon membrane (Amersham) cut to the size of the gel was pre-soaked in water, 10x SSC and then placed on the gel and the air bubbles removed. Three pieces of Whatman 3MM paper cut to gel size were soaked in 10x SSC and placed on top of the membrane followed by paper towels (~10cm thick). A weight (Sigma catalogue x2) was placed on top of the paper towels and the gel blotted overnight (normal samples) or for 48 hours (PFGE samples). The membrane was removed after blotting, rinsed in 2x SSC and baked at 120^{o}C for 30 minutes. For PFGE samples, the gel was re-stained in 1μg/ml ethidium bromide to determine the efficiency of nucleic acid transfer.

2.1.4.2. Dry Blotting

Dry blotting and alkali blotting were used routinely for the transfer of plasmid DNA and smaller molecular weight genomic DNA.

The gel was denatured (0.5M NaOH; 1.5MNaCl) for 30 minutes and neutralised (1.5M NaCl; 0.5M Tris.HCl pH7.4) for 30 minutes. The gel was placed

face down on Saran wrap and Hybond N^+ membrane (pre-soaked in water then 20xSSC) was placed on to the gel followed by 3 pieces of Whatman 3MM paper (soaked in 20xSSC). Paper towels and a weight was placed on top. After overnight blotting the membrane was rinsed in 2xSSC and baked at 120°C for 30 minutes.

2.1.4 3. Alkali Blotting

The gel was denatured in 0.4M NaOH for 20 minutes and dry blotted as above with the Hybond N+ membrane and the Whatman 3MM soaked in 0.4M NaOH

2.1.4.4. Dot Blotting

Dot blotting was used for the routine screening of tail genomic DNA for genotyping purposes. All dot blotting was done by Dianne Peddie using the following one tube method for preparing and blotting genomic DNA from tail biopsies.

Tail biopsies were digested overnight in tail lysis buffer (100mM Tris.HCl pH8.5; 5mM EDTA; 0.2% SDS; 200mM NaCl; 100µg/ml Proteinase K added fresh). While the lysate was still warm, 0.1 ml 5M NaCl was added and vortexed at high speed for 5-10 sec. 0.5 mls of chloroform was added and vortexed again for 5-10 seconds. Samples were spun in a microfuge for 5 minutes. From the aqueous phase, 50 µl was transferred to a 96-well plate and the DNA denatured by adding 150 µl of 0.53 M NaOH and incubating at 37°C for 30 minutes. The dot blot apparatus (Biorad 96-well) was prepared by cutting a piece of Hybond N^+ membrane and a piece of Whatman 3MM paper to fit the apparatus. The membrane was pre-wet in sterile water then soaked in 0.4 M NaOH for 10 minutes. The Whatman paper was pre-wet in 0.4 M NaOH and placed underneath the membrane on the apparatus. Samples were applied to the dot blot apparatus and left for 30 minutes before applying a gentle vacuum to draw through the samples. The membrane was removed, washed with 30mM NaP buffer/0.1% SDS buffer and baked at 120°C for 30 minutes.

2.1.4.5. Northern Blotting

10- 15µg of total RNA was added to 3 volumes of sample buffer (66% deionised formamide; 22% formaldehyde; 1.2xMOPS buffer) and denatured at 70°C for 10 minutes. Samples were cooled on ice and 2µl of 10x RNA loading dye added (50% glycerol; 1mM EDTA; 0.4% bromophenol blue; 0.4% xylene cyanol; 2µg ethidium bromide) the samples were separated on a 1% denaturing agarose gel prepared in 1xMOPS buffer (20mM MOPS; 4mM sodium acetate; 1mM EDTA- pH7.0) and 17.5% formaldehyde (overnight at 25V or 5-6 hours at 70V). Molecular size standards were used (1µg of the 0.24- 9.5Kb RNA ladder - GibcoBRL). After electrophoresis, the RNA quality was determined by the examining the 28s and 18s rRNA bands. If necessary, the gel was stained with EtBr for 30 minutes and destained overnight. The gel was denatured in 50mM NaOH for 30 minutes followed by neutralisation in buffer (1.5M NaCl; Tris.HCl pH 8.0; 1mM EDTA) for 30 minutes. Hybond N^+ membrane was soaked in water and then 20xSSC and the RNA transferred using an overnight capillary dry blot (Section 2.1.4.2.).

2.1.5. Radiolabelling Probes

2.1.5.1. Random Priming Probes

DNA fragments were gel purified as described in Section 2.1.1.1. and the concentration determined by visualisation on an agarose gel. Random priming was performed using the 'High Prime' kit (Boehringer) according to manufacturers instructions.

25-50ng of DNA was added to sterile water to a volume of 12µl, denatured at 100°C for 10 minutes and snap cooled on ice for 5 minutes. To the denatured DNA was added 4µl of high prime mix (1unit/ml Klenow polymerase; 0.125mM dATP; 0.125mM dGTP; 0.125mM; 0.125mM dTTP; 50% v/v glycerol) and 4µl of [α^{32}P]dCTP (3000Ci/mMol, Amersham) and the reaction incubated at 37°C for 10 minutes. The reaction was stopped by adding 1µl 0.5M EDTA pH8.0 and 79ml of

dH_2O and unincorporated nucleotides were removed by centrifugation of the reaction through a G-50 Sephadex column at 2000 x g. for 2 minutes. An aliquot of 2µl from the labelled probe was used to determine [α^{32}P]dCTP incorporation. Prior to hybridisation, the probe fragments were denatured at 100°C for 10 minutes, snap cooled on ice and added directly into the hybridisation mix.

2.1.5.2. End-labelling Oligonucleotide Probes

In a screw cap tube on ice was added 50 ng of oligonucleotide, 2.5µl 10x kinase buffer (Boehringer), 5µl γ-^{32}P-ATP (Amersham), 1µl polynucleotide kinase (Boehringer) and sterile water to a final volume of 25µl. The reaction was incubated at 37°C for 1 hour Unincorporated nucleotides were removed by spinning the reaction through Sephadex-G50 as for random labelled probes. The labelled oligonucleotide was added directly into the hybridisation solution.

2.1.5.3. Table of Probes

Listed below are the probes used in the analysis of the I114 gene trap integration.

Probe	Plasmid	Digest	Size (Kb)
en2-exon	pT1-ATG*	BamHI	0.5
en2-intron	pT1-ATG*	BamHI / HindIII	1.0
lacZ	pT1-ATG*	EcoRI / ClaI	2.2
Group I (LA-8)	pLA-8*	KpnI / XbaI	0.2
Group II (L-69)	pL-69*	EcoRI	0.2
1Bα-fH	p1Bα–2†	HindIII	1.1
1Bα-fXh	p1Bα–2†	XhoI	0.3
4Aα-fP	p4Aα-1†	PstI	0.5
GC7.fE	pGC7†	EcoRI	0.1
GC7.fP	pGC7†	PstI	1.4

GC10.fXb	pGC10†	XbaI	0.35
439-a23(fP).fXh	p439-a23.fP*	XhoI	0.5

UBT-1 oligonucleotide probe (Group I). Stock number 9.

5'-CAACAGCGAGGAAGAGGAGGACGACGACGACGAGGAAGAGGA-3'

(† Appendix III; * Appendix IV)

2.1.6. Hybridisation Conditions

All hybridisations and washes were performed in Techne hybridisation bottles rotating in a Techne HB-1 oven using approximately 10mls of hybridisation buffer per filter. Hybridisation times varied from 3-hour (for high DNA copy number blots e.g. plasmids and PCR products) to overnight (for single copy DNA blots e.g. genomic DNA and RNA blots). Prehybridisation, hybridisation and washes were carried out at 65°C for DNA blots using double stranded probes, 42°C for DNA blots using oligonucleotide probes and 60°C for RNA blots.

Filters were prehybridised for 1-2 hours in hybridisation buffer. For the hybridisation of genomic DNA blots, 0.5M Na_2HPO_4 (pH7.1); 15% deionised formamide (Sigma); 7% SDS; 1mM EDTA; 1% w/v BSA (fatty acid free - Sigma) was used. Church and Gilbert buffer (0.5M Na_2HPO_4 (pH7.1); 7% SDS; 1mM EDTA; Church and Gilbert, 1984) was used for the hybridisation of random primed probes and oligonucleotide probes to high copy number blots including analysis of RACE-PCR and RT-PCR products, PAC clone digests, plasmid digests, secondary screening of cDNA libraries. Northern blots were hybridised in 350mM Na_2HPO_4 (pH7.1); 30% deionised formamide (Sigma); 7%SDS; 1mM EDTA; 1% BSA w/v (fatty acid free, Sigma). Radiolabelled probe was denatured at 100°C for 10minutes, snap cooled and added to fresh hybridisation buffer (~10^6c.p.m./ml). Following hybridisation, blots were rinsed for 3 x 20 minutes in wash buffer (150mM Na_2HPO_4 (pH7.1); 0.1% SDS). After washing, the blots were wrapped in Saran wrap and exposed to autoradiographic film at -70°C for an appropriate length of time (1-7 days) for a signal

to appear. Blots were stripped of hybridised probe by pouring on a solution of boiling 0.5% SDS and allowing to cool to room temperature.

2.1.7. DNA Sequencing

The manipulation of raw sequence data (e.g. contiguous sequence assembly and conceptual translation) was performed using DNAstar Lasergene 1.60 (Lasergene, London, UK).

2.1.7.1. Manual Sequencing

Sequencing of double stranded plasmids (Sanger *et al.*, 1977) was carried out using the Sequenase Version 2.0 DNA Sequencing Kit (USB-Amersham) as per manufacturers instructions with the following modifications. The plasmid was alkali denatured in the presence of the sequencing primer to improve primer annealing and DMSO is added to the reaction to restrict annealing of the template strands (modified from Winship, 1989).

1-5µg of plasmid DNA was added to 1-5pmol of sequencing primer (Section 2.1.7.4.) to a volume of 16µl withT.E. 4µl of NaOH was added and the plasmid denatured for 5 minutes. The plasmid/primer mix was precipitated with 4µl of 2.5M ammonium acetate (pH4.6) and 55µl of 100% ethanol. The samples were spun at 13,000 x g in a microcentrifuge for 10 minutes, pellet washed in 70% ethanol and resuspended in 12.5µl of reaction buffer containing 10% DMSO. The reaction was subsequently carried out as manufacturers instructions except for the addition of DMSO to the termination mix to a final concentration of 10%.

2.1.7.2. Automated Cycle Sequencing

Automated cycle sequencing was performed using the Perkin-Elmer Taq DyeDeoxy Terminator Cycle Sequencing Kit (Applied Biosystems). The cycle sequencing reaction is a modificaton of the dideoxy-termination method of Sanger *et al.*,1977. The four 2',3'-dideoxynucleoside5'triphophates (ddNTPs) are covalently

linked to different fluorescent dyes allowing the sequencing reaction to be carried out in a single tube.

Plasmid DNA was prepared by PEG precipitation or Qiagen extraction. The DyeDeoxy terminator cycle sequencing reactions were carried out in accordance with manufacturers instructions. To conserve reagents, half the amounts were used (200-500ng of plasmid DNA, 1.6pM of primer (Section 2.1.7.4.), 4µl terminator ready reaction mix and sterile water to 10µl volume). The reactions were overlaid with 40µl of mineral oil and subjected to 25 PCR cycles of 96°C for 30 secs; 50°C for 15 secs; 60°C 4 mins. The reactions were precipitated with 1µl 3M sodium acetate (pH4.6) and 2 volumes of 100% ethanol on ice for 15 minutes followed by centrifugation at 13,000 x g for 25 minutes. The pellet was washed in 70% ethanol, air dried and resuspended in 4µl of loading buffer (5mM EDTA pH8.0; 10mg/ml Blue dextran in deionised formamide). The sequencing reactions were denatured at 95°C for 2 minutes and 2µl run on a denaturing polyacrylamide gel (7M urea; 5% acrylamide (29:1 Biorad); 1xTBE; 0.06% ammonium persulphate; 15µl TEMED per 50ml mix) in 1xTBE on the ABI PRISM 377 DNA Sequencer.

ABIPrism™377 Sequencing Software 2.1.1 (Applied Biosystems) was used to analyse the sequencing reactions. Sequence traces were examined by eye to clear any ambiguous sequence. For example, a G nucleotide after an A nucleotide often produced a low signal intensity which resulted in inaccurate base calling. Routinely, each sequencing run produced 500-600 bases.

2.1.7.3. Direct Sequencing of RACE-PCR Products

The following procedure was used to sequence 5'RACE-PCR products (Protocol 1) which had been amplified with the biotinylated primer 105 (Section 2.1.8.1.)

(a) Purification and denaturation of RACE-PCR products

20 µl (200 µg) of Streptavidin coated beads (Dynabeads M-280) were prepared for each template to be sequenced. The beads were immobilised using a magnetic tube

holder (Promega) and the supernatant removed. The beads were resuspended in 20 µl of 1x B&W buffer and mixed gently (2x B&W buffer was made;10mM Tris-HCl pH 7.5, 1.0 mM EDTA, 2.0 M NaCl). Supernatant was removed after immobilising the beads on the magnet and then the beads were resuspended in 40 µl of 2x B&W buffer. Biotinylated RACE-PCR products (5-40µl) were diluted to 40µl with sterile water and mixed with 40µl of washed beads (in 2xB&W). Samples were incubated for 15 minutes at room temperature with frequent mixing to keep beads suspended. Using a magnet, the supernatant was removed from the beads which were then washed with 40µl of 1x B&W (samples stable for several weeks at 4°C). The supernatant was removed, the beads resuspended in 8µl of fresh 0.1MNaOH and the sample incubated for 10 minutes at room temperature. The supernatant was removed and the beads washed once with 50µl 0.1MNaOH, once with 40µl 1xB&W buffer and then once with 50µl T.E. The beads were resuspended in 25µl of sterile water.

(b) End labelling of sequencing primer

10 pmol of sequencing primer 98 (Section 2.1.7.4.) was added to 2 µl 10X polynucleotide kinase buffer (Boehringer-Mannheim); 50 µCi [g-^{32}P] ATP (3000-6000 Ci/mM); 1µl (10 units) polynucleotide kinase (Boehringer-Mannheim) to a total volume of 20 µl with sterile water. Samples were incubated at 37°C for 10 minutes, heat inactivated at 95°C for 2 minutes and then stored at -20°C.

(c) Sequencing reactions

Sequencing reactions were carried out as per manufacturers instructions using the Amersham ThermoSequenase Kit (US 78500). To a 0.5ml PCR tube on ice was added 25µl of streptavidin bound RACE-PCR products; 2µl 10x ThermoSequenase reaction buffer; 1µl [^{32}P] - end labelled primer (0.5 pmole); 2µl ThermoSequenase. 6µl of the reaction mix was added to each of four 0.5ml PCR tubes containing 2µl of each termination mix (ddATP, ddCTP, ddGTP, ddGTP). The reactions were overlaid with 40µl mineral oil and placed in a preheated (95oC) thermal cycler for 2 minutes followed by 25 cycles of 95°C for 30 secs; 60°C for 30 secs; 72°C for 1 min. When

the reaction was complete 4µl of STOP solution was added, the samples heated to 80°C for 3 minutes and the samples visualised by electrophoresis on a 6%polyacryamide gel.

2.1.7.4. Sequencing Primers

Primer (Stock No.)	Sequence 5' - 3'
Vector	
T3 (20)	AATTAACCCTCAACTAAAGGG
T7 (21)	GTAATACGACTCACTATAGGGC
SP6 (62)	AGCTATTTAGGTGACACTATAG
Endogenous cDNA	
SANK-2 (69)	GAACTGGCAGCAAATTGG
SANK-3 (70)	CACCACCTTCACATGACC
SANK-4 (71)	GAAGGCTGGCTTTGGCTG
SANK-5 (72)	GTAAACTGGATGGCATGG
1.5kb PstI PAC fragment (439-a23.fP)	
SPAC-1 (73)	TTAGTATCGCGAGACGAG
SPAC-2 (74)	TGCGGTGGCTTCTTCTTC
Direct Sequencing	
98	AGCAGTGAAGGCTGTGC

2.1.8. Polymerase Chain Reaction (PCR)

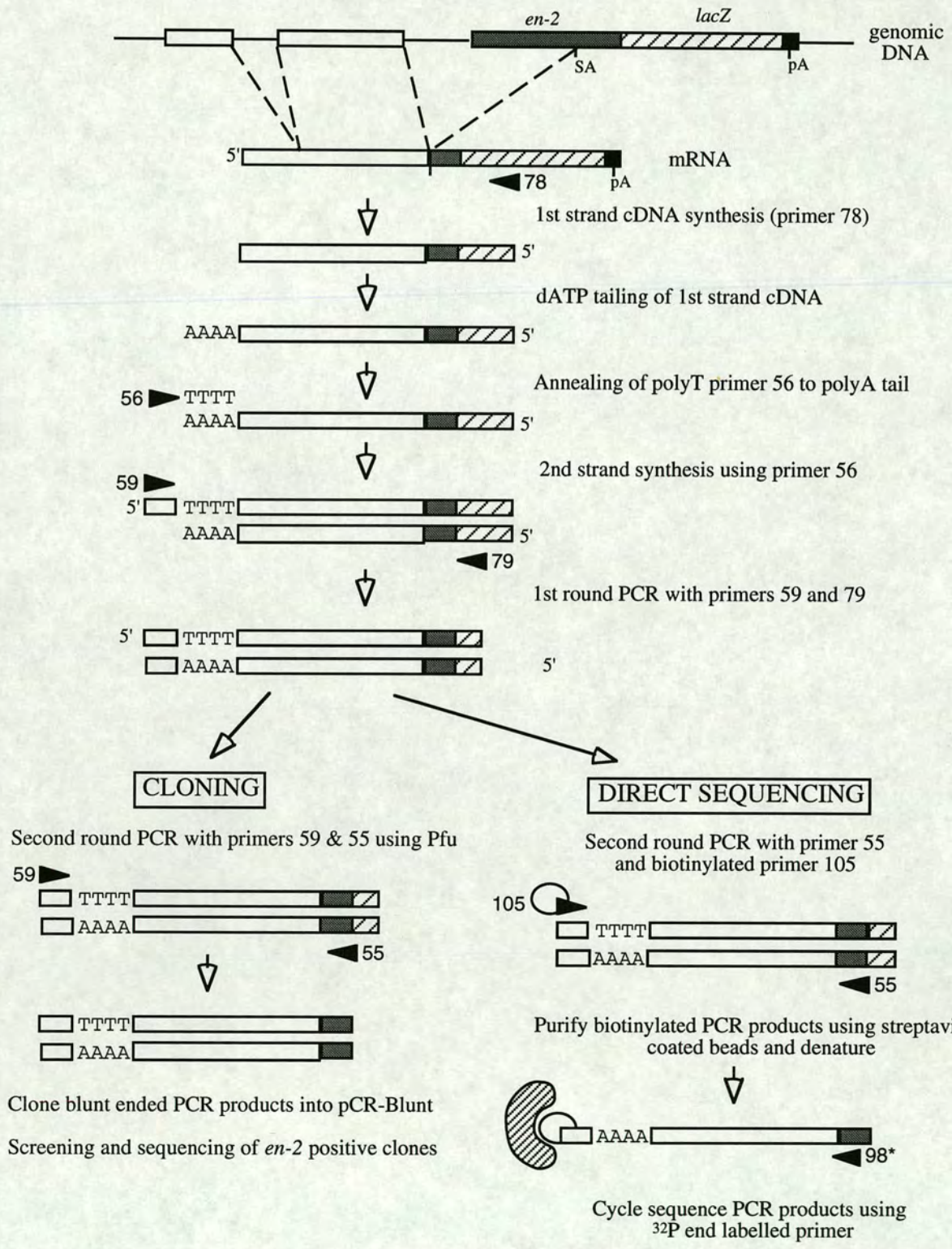
All PCR reactions were carried out using the Hybaid Omnigene Thermocycler. dNTP's were supplied by Boehringer.

2.1.8.1. Rapid Amplification of cDNA Ends-PCR (5'RACE-PCR)

The 5'RACE protocol was initially described by Frohman *et al.*, 1988 with the modifications to that protocol as described by Townley *et al.*, 1997 essentially providing both the protocols described below.

Protocols 1 and 2 are essentially the same except for the use of different primers (Section 2.1.8.1. Protocol 1(g); Section 2.1.8.1. Protocol 2(g)) and several procedural differences. Protocol 1 is described in detail with the corresponding primers used in Protocol 2 given in parentheses. Procedural differences between Protocol 1 and Protocol 2 are listed after protocol 1.

(a) **Figure 2.1:** Overview of 5'RACE-PCR (Protocol 1)



Protocol 1 (Figure 2.1)

(a) First Strand cDNA Synthesis

To a 0.5ml PCR tube on ice was added 10 µg of total RNA; 10ng primer 78 (GSP1-protocol 2) with DEPC treated water added to a volume of 12µl. Samples were denatured at 70°C for 5 minutes followed by snap cooling on ice. The contents were collected by brief centrifugation. To the RNA was added 4µl 5x 1st strand buffer (520mM Tris-HCl pH8.3; 375mM KCl; 15mM $MgCl_2$ - GibcoBRL); 2µl 10mM DTT; 1µl 10mM dNTP. After heating at 37°C for 2 minutes, 1µl (200units) of Superscript II (GibcoBRL) was added to a total volume of 20µl and the samples incubated at 37°C for 1 hour. 1M NaOH was added to 0.1M final concentration (2.2µl) and the reactions incubated at 65°C for 20 minutes to hydrolyse RNA. Samples were neutralised with 2.2µl 1M HCl. The first strand cDNA reaction was then dialysed against T.E. for 4 hours using a 0.025µm microdialysis filter (Millipore) floating in a petri dish with ~50mls of T.E. Approximately, 5-10µl was recovered after dialysis. Sterile dH_20 was added to 20µl.

(b) Poly A Tail Addition

To 20µl of 1st strand reaction was added 6µl 6xTdT buffer (0.1M Potassium cacodylate (pH7.2); 10mM $CoCl_2$; 1mM DTT); 2µl 2mM dATP. Samples were incubated at 37°C for 2 minutes and then 2µl (20units) of terminal transferase (TdT-GibcoBRL) was added and the reaction incubated at 37°C for 5 minutes. The reaction was heat inactivated at 70°C for 2 minutes and the samples collected by centrifugation

(c) 2nd Strand Synthesis (not in protocol 2)

To 15µl of the tailing reaction was added 2µl 10x restriction buffer M (Boehringer); 1µl 10mM dNTPs; 10ng primer 56; 1µl (2units) Klenow (Boehringer) The samples were incubated at room temperature for 30 minutes followed by 30 minutes at 37°C. The reaction was heat inactivated for 5 minutes 70°C. The double stranded cDNA sample was microdialysed on 0.1µm microdialysis filter against T.E. as above. dH_20 was added to 37µl.

(d) 1st Round PCR Amplification

To 37µl of the double stranded cDNA was added 5µl 10x PCR buffer (50mM KCl/10mM Tris (pH8.3)); 4µl 25mM $MgCl_2$; 1µl 10mM dNTP's; 100ng primer 59 (anchor- protocol 2); 100ng primer 79 (GSP 2- protocol 2); 1µl (5 units) Taq DNA polymerase (Promega). The reaction overlaid with 50µl of mineral oil. The PCR amplification was performed on a thermocycler through 30 cycles of 94°C for 1min 30 secs; 60°C for 1min 30 secs; 72°C for 3 mins. PCR products were dialysed against T.E. on 0.1µm filters as above. Approximately 15µl was recovered after dialysis. dH_20 was added to 40µl.

(e) 2nd Round PCR Amplification

If the products were to be taken through the direct sequencing protocol (Section 2.1.7.3.) then 5µl of 1st round PCR products; 5µl 10x PCR buffer; 4µl 25mM $MgCl_2$; 1µl 10mM dNTPs; 100ng primer 105; 100ng primer 55; 1µl Taq DNA polymerase (Promega) were added to a 0.5ml PCR tube.

If the PCR products were to be cloned (Section 2.1.8.1. Protocol 1(f)), the nested second round PCR amplification reactions were performed as follows using Pfu polymerase (GibcoBRL) to produce blunt ended PCR products for cloning using the Zero Blunt™ PCR cloning kit (Invitrogen). 5µl 1st round PCR products; 5µl 10x Pfu buffer (Stratagene); 1µl 10mM dNTPs; 100ng primer 59 (anchor- protocol 2); 100ng primer 55 (55- protocol 2); 1µl Pfu polymerase (Stratagene).

Both reaction volumes were made up to 50µl with sterile water. The same PCR cycle conditions and microdialysis conditions were carried out as for the 1st round PCR reactions. To assess the specificity of the PCR amplifications, 5µl of the 1st and 2nd round PCR products were ran on a agarose gel, Southern blotted and hybridised to the *en-2* exon probe.

(f) Cloning of PCR Products

The Pfu amplified 2nd round PCR products were run on a 1.0% LMP agarose gel and the amplification products between 1.0 and 0.5kb in size were purified from the gel by phenol extraction and precipitation (Section 2.1.1.1.). Half of the remaining

PCR products were subsequently cloned into pCR®-Blunt using the Zero Blunt™ cloning procedure (Section 2.1.1.).

(g) Primers used in Protocol 1

Primer (Stock No.)	Sequence 5' -> 3'(homology)
56	GGTTGTGAGCTCTTCTAGATGGTTTTTTTTTTTTTTTT (anchor)
59 (19)	GGTTGTGAGCTCTTCTAGATGG (anchor)
105 (17)	GGTTGTGAGCTCTTCTAGATGG (anchor, 5'biotinylated)
78 (44)	TAATGGGATAGGTTACG (*LacZ*)
79	AGTATCGGCCTCAGGAAGATCG (*LacZ*)
98	AGCAGTGAAGGCTGTGC (*en-2*)

Protocol 2

Is essentially the same as protocol 1 except for the following differences.

(b) Poly C Tail Addition

To 20µl of 1st strand reaction was added 6µl 6xTdT buffer; 2µl 2mM **dCTP** and the samples incubated at 37°C for 2 minutes. 2µl (20units) of terminal transferase (TdT- GibcoBRL) was added and the reaction incubated at 37°C for 5 minutes followed by heat inactivation at 70°C for 2 minutes. Samples were collected by centrifugation. The PolyC tailed cDNA products were used directly in the 1st round PCR amplification using the polyG anchor primer. No second strand synthesis was necessary.

(e) 2nd Round PCR

As for Protocol 1 for direct sequencing using Taq DNA polymerase except 100ng of polyG anchor primer used.

(f) Cloning of PCR Products

2nd round PCR products were gel purified, digested with SpeI and KpnI and directionally cloned into a SpeI/KpnI digested pBluescript II KS- (Section 2.1.1.2.).

(g) Primers used in Protocol 2

Primer (Stock No.)	Sequence 5' -> 3'(homology)
Anchor (4)	GGCCACGCGTCGACTAGTACGGGIIGGGIIGGGIIG
GSP1 (1)	GCAAGGCGATTAAGTTGGGT (*LacZ*)
GSP2 (2)	CCGTCGACTCTGGCGCCGCT (*en-2*)
55 (40)	TGCTCTGTCAGGTACCTGTTG (*en-2*)

Underlined nucleotides correspond to restriction sites SpeI (anchor) and KpnI (55) used for cloning RACE-PCR products.

2.1.8.2. Reverse transcriptase PCR (RT-PCR)

(a) cDNA Synthesis

1mg of total RNA was made up to a volume of 13µl with DEPC treated water in a 0.5ml PCR tube. Samples were incubated at 70°C for 5 minutes to denature any RNA secondary structure, chilled on ice for 2 minutes and the samples collected at the bottom of the tube by centrifugation. To each sample (on ice) was added 2µl 10x PCR buffer (Promega); 2µl 25mM $MgCl_2$; 1µl (0.125 units) random hexamers (Boehringer); 2µl 10mM dNTPs; 1.0µl (200units) MMLV reverse transcriptase (GibcoBRL)and DEPC treated water to a final volume of 20µl. Samples were left at 25°C for 10 minutes, incubated at 42°C for 1 hour then denatured at 95°C for 10 minutes.

(b) PCR

To 4µl of the first strand cDNA reaction was added: 1.6µl 10x PCR buffer; 1.6µl 25mM $MgCl_2$; 100ng primer x; 100ng primer y; 0.2µl (1 unit) Taq DNA polymerase (Promega) with water added to 20µl final volume. Samples were overlayed with 30µl of mineral oil (Sigma) and PCR amplification carried out using 30 cycles of 96°C for 5 secs; 53°C for 15 secs; 72°C for 1 min.

(c) Primers used in RT-PCR

Primer (Stock No.)	Sequence 5' - 3' (homology)
LST-1 (38)	GGTAGTTTTCTGTCAGTGG (Group II)
LST-2 (39)	CCCCTCCACAACTGCTCC (Group II)
RTANK-2 (66)	CTGTGCTCTCAGCTCTCATC (endogenous cDNA)
RTANK-3 (67)	GAAGACTGTGCATTGACATC (endogenous cDNA)
HPRT3' (22)	GCTGGTGAAAAGGACCTCT (HPRT)
HPRT5' (23)	CACAGGACTAGAACACCTGC (HPRT)

2.1.9. RNase Protection Assay

The following plasmids were used for the generation of antisense riboprobes. Plasmids were linearised by restriction digest using the specified enzyme and the riboprobe generated using the appropriate RNA polymerase site.

Riboprobe	**Plasmid**	**Digest**	**Polymerase**
Group I (LA-8)	pLA-8*	XbaI	T3
Group II (L-69)	pL-69*	HindIII	T7
5Aα.fP(150)	p5Aa.fP(150)*	XbaI	T3
GAPdH	GAPdH*	AccI	SP6

* Appendix IV

The linearised plasmid was gel purified (Section 2.1.1.1.) and the concentration determined after separation of the purified fragment on an agarose gel.

(a) Riboprobe Synthesis

In a screw cap tube at room temperature was added 2µl (500ng) of template DNA; 0.75µl 200mM DTT; 0.75µl 2mg/ml BSA; 2.25µl 3.3mM ATP/UTP/GTP; 0.5µl (20units) RNase inhibitor (GibcoBRL); 6.25µl [$a^{32}P$] CTP (250mCi);1.5µl 10x transcription buffer (40mM Tris.HCl pH8.0; 6mM $MgCl_2$; 10mM DTT; 2mM spermidine); 1.0µl RNA polymerase (T3, T7 or SP6- Boehringer). For the GAPdH loading control, 0.5µl [$a^{32}P$] CTP and 5.75µl unlabelled 1mM CTP for GAPdH was

added in place of the 6.25µl [$a^{32}P$] CTP. The *in vitro* transcription reaction was incubated at 37°C for 1 hour and then treated with 2µl (20units) RNase-free DNase at 37°C for 15 minutes to remove the DNA template. The reaction volume was made up to 50µl with sterile water. To remove protein and unincorporated nucleotides, the probe was either extracted once with an equal volume of phenol/chloroform and spun through a Sephadex G-50 column or centrifuged through a SpinX column (Sigma Costar) loaded with Sephadex G-50 at 13,000 x g for 5 seconds. The probe was either directly gel purified from this stage or stored overnight at -80°C.

(b) Gel Purification of Riboprobe

The riboprobe was mixed with loading dye (95% formamide; 20mM EDTA; 0.05% bromophenol blue; 0.05% xylene cyanol FF) at a ration of 3:2, denatured at 80°C for 2 minutes and run on a 6% denaturing polyacrylamide sequencing gel (7M urea; 5.7% acrylamide; 0.3% bisacrylamide; 1xTBE; 0.06% ammonium persulphate; 35µl TEMED per 100ml mix) in 1xTBE for ~1.5hours at 60W. After the electrophoresis the wet gel on the glass plate was wrapped in Saran wrap and exposed to autoradiographic film for 5 minutes (the film was aligned to the top left of the glass plate). The developed autoradiograph was aligned to the gel and the area of the gel containing the largest probe band excised using a scalpel. The gel was re-exposed to ensure that the correct region of the gel had been isolated. The gel slice was incubated in 0.5ml of probe elution buffer (0.5M ammonium acetate; 1mM EDTA; 0.2% SDS) at 37°C for 2 hours in a screw cap tube with vigorous shaking to elute the probe. The gel was pelleted by centrifugation at 13,000 x g for 5 minutes and the supernatant containing the probe removed. To measure the incorporation of each probe, 2µl of riboprobe was added to 1ml of scintillation fluid (Ultima Gold).

(c) Hybridisation

10µg target RNA, 3.5 x 10^5cpm of eluted riboprobe and 5 x 10^4cpm of GAPdH loading control riboprobe in a total volume 100µl was precipitated with 0.3M sodium acetate, 20mg glycogen (optional) and 2 volumes of 100% ethanol at -80°C for 30 minutes. The RNA was pelleted by centrifugation at 13,000 x g for 10 minutes and

then washed with 70% ethanol and air dried. The RNA pellet was resuspended in 30µl of hybridisation mix (6µl 5x hybridisation buffer (2M NaCl; 200mM PIPES pH6.4; 10mM EDTA); 24µl deionised formamide). Hybridisation reactions were then denatured at 85°C for 15 minutes and hybridised overnight at 55°C.

(d) RNase Digestion

Digestion buffer was made up with 60µl 1MTris.HCl pH7.5; 60µl 0.5M EDTA; 360µl 5M NaCl; 5.25ml sterile water; 9.6µl (25mg/ml) RNaseA; 2.1µl (2000 units) RNase T1 (GibcoBRL). 350µl of digestion buffer was added to the hybridisation reaction and incubated at 37°C for 30 minutes to remove unhybridised RNA. The samples were then incubated for a further 10 minutes with 10µl of 20%SDS and 5µl of 10µg/ml proteinase K to stop the digestion. The reactions were then extracted with phenol/chloroform and precipitated with 1ml 100% ethanol (5mg tRNA as a carrier-optional) at -80°C for 30 minutes. RNA was pelleted by centrifugation at 13,000 x g for 10 minutes, air dried and resuspended in 4µl of loading dye. Samples were then denatured briefly at 95°C and separated on a 6% polyacrylamide gel at 60W for 2-3 hours depending on the predicted size of the digestion products. The gel was dried before exposing to autoradiographic film at -80°C.

2.1.10. cDNA Library Screening

The protocol for screening λZAP®II (Stratagene) libraries is essentially as manufacturers instructions. The λZAPII vector contains the pBluescript SK(-) phagemid into which the cDNAs are subcloned. Screening of the library to single colonies is performed in λZAPII. Using the *in vivo* excision protocol, pBluescript SK (-) can then be isolated as a plasmid.

(a) Maintenance of Bacterial Strains

XL1 Blue MRF' is used to propagate λZAPII and is maintained as a glycerol stock and bacterial streak using LB + tetracycline (12.5µg/ml). For incubating λZAPII, XL1 Blue MRF' is plated on LB agar + 0.2% maltose + 10mM $MgSO_4$.

SOLR™ strain is used in the *in vivo* excision protocol to recover excised phagemids and is maintained as a glycerol stock and bacterial streak using LB + kanamycin (50µg/ml). For incubating the excised phagemid, SOLR™ is plated on LB agar with no supplements.

Plating cultures were inoculated from a fresh LB plate (with appropriate antibiotic) of cells using a single colony. Cultures were either grown overnight at 30°C in 50 ml LB (+/- supplements) or grown overnight at 37oC in 10mls LB (+ supplements) and added to fresh 50mls LB (+/- supplements) and grown for 2-3 hours. The cultures were spun at 2000 x g for 10 minutes in a Beckman-GPR benchtop centrifuge and gently resuspended in 10 mM $MgSO_4$ to the appropriate OD_{600} (0.5 for XL1-Blue and 1.0 for SOLR)

(b) Titering Procedure

Titers of phage libraries and selected phage clones were determined by making several serial dilutions of the phage in SM buffer (100mM NaCl; 10mM $MgSO_4$; 50mM Tris.HCl pH7.5; 0.01% gelatin). Up to 100µl of the diluted phage was added to 200µl OD_{600}= 0.5 XL1-Blue in a Falcon 2059 tube and the phage were allowed to adsorb for 15 minutes at 37°C. 3.5mls of molten (45°C) top agarose (0.7% agarose in 10mM $MgSO_4$) was added and immediately poured onto pre-warmed (37°C) 9cm LB plates. The top agarose was allowed to set and the plates incubated inverted at 37°C overnight. Counting the number of plaques determined the plaque forming units (pfu) per ml concentration of the original phage stock.

(c) Library Plating and Plaque Lifts

Phage stocks were mixed with plating culture OD_{600}= 0.5 XL1-Blue, incubated at 37°C for 15 minutes and then moulten top agarose(45°C) added. The following proportions were used depending on the size of bacterial plate:

plate size	maximum no. phage(pfu/ml)	volume top agarose
9cm	~5,000	3.5ml
15cm	20-30,000	7ml
20 x 20cm	150-200,000	20-25ml

The top agarose was allowed to set and the plates incubated at 37°C overnight. Plates were left at 4°C for 1-2 hours to allow agarose to harden before performing plaque lifts. Hybond N^+ nylon membrane (Amersham) was placed on to the agarose (avoiding bubbles) for 1 minute and asymmetric orientation marks made on the filter with a needle. A duplicate lift was taken by placing a fresh piece of membrane on the agarose for 2 minutes (using the same orientation marks). Membranes were placed face up on Whatman 3MM paper soaked in: denaturing solution (1.5M NaCl; 0.5M NaOH) for 2 minutes; neutralising solution (0.5M Tris.HCl pH8.0; 1.5M NaCl) for 3 minutes; neutralising solution for 3 minutes and then rinsed in 2xSSC. The membranes were then air dried briefly and baked at 120°C for 30 minutes.

(d) Library Screening

In the primary screening a total of $1x10^6$ pfu were screened from a random-primed D3 ES cell cDNA library cloned into λZAP II (gift from Hitoshi Niwa). Phage were plated onto 20cm x 20cm LB plates at the above proportions.

Filters were individually soaked in prehybridisation buffer (5xDenhardt's; 6xSSC; 0.5% Sarkosyl; 100mg/ml denatured herring sperm) and then incubated at 65°C for 2-3 hours in a sealed tupperware box on a shaking platform. The prehybridisation buffer was replaced with an equal volume of hybridisation buffer (5xDenhardt's; 6xSSC; 0.5% Sarkosyl; 100mg/ml denatured herring sperm; 100mg/ml dextran sulphate). ^{32}P-labelled probe was added to the hybridisation buffer at 10^6cpm/ml and incubated at 65°C overnight. After hybridisation, filters were washed in 2xSSC; 0.1%SDS four times for 30 minutes at 65°C and subsequently exposed to autoradiographic film at -80°C for 1-5 days depending on the signal intensity. From the primary screen, plaques with a duplicate positive signal were cored with the wide end of a sterile blue Gilson tip, placed in 500µl of SM buffer and the phage eluted by shaking or overnight incubation. Secondary screening was performed by plating serial dilutions of each primary plaque onto 9cm agar plates and the same protocol followed as for the primary screen. Hybridisation from the secondary screen onwards was

performed in Church and Gilbert solution (Section 2.1.6). This procedure was repeated until a single, well isolated positive plaque could be purified.

(e) *In vivo* Excision

The pBluescript SK(-) plasmid carrying the cDNA insert was excised from λZAPII using the Ex Assist helper phage (Stratagene). In λZAPII, pBluescript is flanked by the replication initiator (I) and terminator (T) signals of bacteriophage f1. Co-infection of λZAPII and f1 into *E.coli* results in the replication of pBluescript and any potential cDNA insert as a single stranded molecule by the co-infected f1 phage proteins. The newly synthesised pBluescript DNA is circularised and packaged as a f1 phage. Subsequently, the pBluescript can be recovered by infecting an F' strain (SOLR) and growing in the presence of Ampicillin. The helper phage carries an amber mutation that prevents its replication in a non suppressing *E.coli* strain such as SOLR. This allows only the excised plasmid to replicate in the host.

In a 50 ml conical tube was added 200µl of OD_{600}= 1.0 XL1-Blue MRF' cells; 100µl of λZAPII phage stock (>$1x10^5$ pfu) from a positive plaque; 1µl of ExAssist helper phage (>$1x10^6$ pfu/ml). The tube was incubated at 37°C for 15 minutes and then 3 mls of LB broth added and incubated at 37°C for 2-2.5 hours with shaking. The cultures were incubated at 70°C for 15 minutes to kill the bacterial and the dead cells pelletted by centrifugation at 4000 x g for 15 minutes. The supernatent contains pBluescript packaged as f1 phage (may be stored at 4°C for up to 2 months). To recover the excised phagemids, 200µl of a plating culture of SOLR cells (OD_{600}= 1.0) was mixed with 1, 10 and 100µl of the supernatent and incubated at 37°C for 15 minutes. 100µl from each tube was plated on LB/Amp plates which were incubated at 37°C overnight. Approximately 5 colonies were selected per individual cDNA clone for plasmid preparation to confirm the presence of a cDNA insert.

2.2. Protein Analysis

The followiing procedure is adapted from Ausubel *et al.* (1998) using the Laemmli system (Laemmli, 1970). SDS-PAGE and immunoblotting was carried out using the Mini-PROTEAN II apparatus (Biorad).

2.2.1. Protein Preparation

Crude protein lysates were prepared by homogenising between 100-500mg of tissue in a ground glass homogeniser in 2mls of lysis buffer (2% SDS/ 50mM Tris pH7.4). The sample was then centrifuged for 5 minutes at 13,000 x g and the supernatent removed. The protein concentration of each sample was determined using the Biorad protein assay kit according to the manufacturers instructions. The OD_{595} of each sample was compared to a protein concentration curve using BSA standards.

2.2.2. SDS-Polyacryamide Gel Electrophoresis (SDS-PAGE)

The separating gel mix was prepared by adding 7.0ml dH_2O; 4.0ml 30% acrylamide mix (29.2% acylamide; 0.8% N1, NI-methylene-bis acrylamide); 3.8ml 1.5M Tris (pH8.8); 150µl 10% SDS; 9µl TEMED (Sigma); 150µl 10% ammonium persulphate (freshly made). Gel plates were cleaned, assembled and the separating gel poured. The gel was overlayed with Butan-2-ol -dH2O saturated (top phase) and allowed to set for 30 minutes. The set gel was then rinsed thoroughly with sterile water to remove any trace of Butan-2-ol

The stacking gel mix was prepared by adding 3.4ml dH_2O; 830µl 30% acrylamide mix; 630µl 1M Tris (pH6.8); 50µl 10% SDS; 5µl TEMED; 50µl 10% APS. The stacking gel was poured on top of the separating gel, the combs were inserted and the gel allowed to set for 10 minutes. 5x Laemmli Sample buffer dye (62.5mM Tris (pH6.8); 12.5% glycerol; 2% SDS; 20mM DTT; 0.0025% bromophenol Blue) was heated to 100^{o}C and an added to 10-20mg protein samples. The protein samples and an aliquot (10µl) prestained SDS-PAGE broad-range molecular weight

markers (Biorad Cat No. 72807A) were denatured for 5 minutes at 100°C. Samples were loaded on the gel and ran at 90mA for 50minutes using Biorad's Mini-Protean II electrophoresis tank in 1x Laemmli running buffer (25mM Tris; 192mM glycine; 1% SDS).

2.2.3. Coomassie Staining

The electrophoresis equipment was disassembled, the gel plates separated, the stacking gel removed and an orientation mark put on the gel The gel was fixed for 15-30 minutes in fix solution (20% methanol; 10% acetic acid) and then Coomassie stain (0.25% w/v Coomassie blue R; 50% methanol; 10% acetic acid) was added for 20 minutes. The gel was destained overnight at room temperature in water.

2.2.4. Immunoblotting

Sponge pads (Biorad), Hybond ECL membrane (Amersham) and the separating gel were equilibrated in protein transfer buffer (25mM Tris; 192mM glycine; 20% methanol) for 10-15 minutes at 4°C (membrane was pre-soaked in dH_2O). Blotting apparatus (Mini-PROTEAN II) was assembled in protein transfer buffer according to manufacturers instructions. The ice block was added to the protein transfer buffer (4°C) and then the gel cassette was added. Blotting was carried out at 100mA for 30 minutes and then 300mA for 30 minutes with stirring. The apparatus was disassembled, the filter removed and rinsed briefly in PBS.

2.2.5. Immunodetection

The membrane was blocked in 10mls 5% milk (Marvel) in TBS-T for 1hr at room temperature (or 4oC O/N). The filter was probed with a polyclonal rabbit anti-mouse α-foetoprotein 1° antibody (ICN Biomedicals Cat. No: 64-561) at 1:1000 dilution in TBST with 0.3% milk, 0.02% sodium azide for a minimum of 1 hour at room temperature. The filter was washed 5 times for 5 minutes each in TBST at room temperature. The filter was then probed with Anti-rabbit Ig- horseradish peroxidase

linked whole antibody (Amersham NA934) at 1:5000 (2µl in 10 mls) for 1hour at room temperature. The filter was washed 5 times for 5 minutes each in TBST at room temperature. Equal volumes (4mls each) of ECL reagents (Amersham) were added to the filter and left for 1 minute. The membrane was quickly exposed to film for 5, 10, 30 and 60 seconds. The ECL system is a luminol based chemiluminescent reaction for the detection of horseradish peroxidase labelled antibodies.

2.3. ES Cell Culture

General methods for the manipulation of ES cells are based on Hogan *et al.*, 1994. All ES cell manipulations were performed in laminar flow sterile hoods using a strict sterile technique which included wiping the hood down and spraying all items entering the hood with 70% industrial methylated spirits (IMS). ES cells were incubated at 7.5%CO_2 at 37°C in a humidified incubator (Heraeus). All solutions were filtered and subsequently tested for sterility and were also prewarmed to 37°C prior to use. ES cells were examined using an inverted microscope (Olympus CK2).

2.3.1. Reagents

The parental R1 ES cell line and the I114 gene trap cell line were maintaied in 1x G-MEM ES cells culture medium containing 15% FCS (1x Glasgow MEM (GibcoBRL); 0.25% sodium bicarbonate (GibcoBRL); 0.1% MEM non-essential amino acids (GibcoBRL); 4mM glutamine (GibcoBRL); 2mM sodium pyruvate (GibcoBRL); 0.1mM 2-mercaptoethanol (Sigma); 15% foetal calf serum (FCS) (Globepharm, Surrey). All flasks were gelatinised (5 minutes with 0.1% gelatin in PBS) prior to addition of ES cells. ES cells were maintained in an undifferentiated state by Differentiation Inhibiting Activity/Leukemia Inhibitory Factor (DIA/LIF). Murine or human DIA/LIF expression plasmids were used to transiently express DIA/LIF in COS-7 cells using the previously described method (Smith, 1991). Serial dilutions of the supernatant were tested on ES cells for their ability to maintain pluripotency. Routinely, 100x the mininum concentration required to keep ES cells undifferentiated was used as the working concentration. All the stock solutions were prepared by Douglas Colby and Derek Rout at the CGR.

2.3.2. Thawing ES Cells

Frozen ES cell vials were taken directly from the liquid nitrogen storage and quickly thawed in a 37°C water bath. The cell suspension was transferred to a

centrifuge tube containing 10ml of prewarmed ES cell medium and immediately centrifuged at 1,200 x g for 3 minutes (Denley BS400 Benctop). The media was aspirated and the pelleted cells were resuspended in 1ml of ES cell medium. The cell suspension was then used to seed a 25cm^2 flask containing 9mls of ES cell medium (+DIA/LIF) and the medium changed after 8 hours of culture to remove remove any dead cells.

2.3.3. Passsage and Expansion of ES Cells

Cultures were monitored every day to assure that ES cells had not grown past confluency. Cells were normally passaged every two days. Culture medium was aspirated off the ES cell culture and the cells rinsed twice with 5mls PBS for a 25cm^2 flask. 1ml of trypsin solution (0.025% trypsin (Gibco); 0.1% chicken serum (Flow Labs); 1.3mM EDTA (Sigma) in PBS) was added to the cells and incubated at 37^{o}C for 1-2 minutes until a single cell suspension was achieved. 9mls of ES cell medium was added to neutralise the trypsin and the cell suspension was centrifuged for 5 minutes at 1,200 x g, excess media was removed and the pellet resuspended in 5mls of fresh culture media. The cell were counted in a haemocytometer and 1x10^6 cells in 10 mls of ES cell medium supplemented with DIA/LIF (100units/ml) were seeded to a 25cm^2 gelatinised flask.

2.3.4. Freezing ES Cells

ES cells were frozen at 1 vial per 25cm^2. Cells were trypsinised into a single cell suspension as above. The cell suspension was centrifuged for 5 minutes at 1,200 x g and the pellet resuspended in 0.5ml of freezing mix (10% dimethyl sulphoxide in ES cell medium) and rapidly aliquotted into Nunc cryotubes on ice. Vials were put immediately at -80^{o}C overnight before being transferred to a liquid nitrogen cell bank for long term storage. Exposure of ES cells to DMSO is kept to a minimum as it is toxic to cells and is an ES cell differentiation agent.

2.3.5. Selection of Retinoic Acid-Responsive Gene Trap Cell Lines

After electroporation of the gene trap vector PT1-ATG (Hill and Wurst, 1993), the R1 ES cells were maintained in G418 for 8 days. G418 resistant colonies were selected, replica plated onto filters (Hill and Wurst, 1993) and the filters treated with either ES cell medium or RA-induction medium (ES cell medium (without LIF) containing 5%FCS and 10^{-6}M all-*trans*-RA - Sigma). After 42 hours, fresh medium was added to the filters and the filters left for an additional 6 hours before assaying the for β- gal activity (Forrester *et al.*, 1996).

2.4. Histology

2.4.1. Maintenance of Animals

All mice were housed and bred within the Centre for Genome Research, Edinburgh, according to the provisions of the animals (Scientific Procedures) Act (UK) 1986. Mice were housed in a stabilised enviroment with a 14 hours light/ 10 hours dark cycle (midpoint 12 O'Clock, midnight). They were provided with a constant supply of water and chow food. As standard practice, litters from natural matings were left with the parents until 3 weeks of age when they were weaned by separating the offspring from their parents. At weaning animals are sexed and tail tips taken for genotyping. At 6 weeks of age the mice can be used for mating.

2.4.2. Preparation of Specimens for Histology

All animals were culled by the schedule 1 method of cervical dislocation.

For the collection of embryos at specific stages of gestation, matings were set up overnight and the females examined for the presence of a vaginal plug the next morning. If a vaginal plug was found, the stage of gestation was marked as 0.5 days post coitus. Pregnant females at the correct stage in gestation were sacrificed and the uterus dissected out in to cold PBS. Subsequently, the embryos were dissected from the uterus and all the extraembryonic tissues (decidua, placenta, Reichert's membrane) removed with the exception of the yolk sac. The age of the embryos up to approximately 10.5d.p.c. was determined from somite number. Foetuses 11d.p.c. and older were decapitated immediately. Adult tissues were dissected free of connective tissues and fat and placed in to ice cold PBS for cryosectioning.

2.4.3. Cryostat Sectioning

Freshly dissected tissues or embryos were embedded in OCT compound (Tissue Tek, Miles Inc, Diagnostic Division, Ellchort, IN, USA), snap frozen in liquid nitrogen and stored at -80^{o}C. Prior to sectioning, OCT mounted tissues were placed in

the cryostat (Cryotomb 650- Anglia Scientific) and allowed to equlibrate to the cutting temperature (-15 to -20°C) for ~1 hour. Sections were typically cut at 10μm and lifted directly onto room temperature TESPA coated slides (see below). The sections were dried onto the slides at room temperature and either stored at -20°C or stained for β-gal activity.

TESPA Coating Slides

Glass microscope slides (Chance Popper) were treated with 10% HCl in 70% ethanol for 10 seconds, rinsed in dH_2O for 10 seconds, dehydrated in 95% ethanol for 10 seconds and baked dry at 150°C for 5 minutes. Once cooled to room temperature the slides were dipped in 2% TESPA (3-aminopropyl-tritheoxysilane; Sigma, A3648) in acetone for 10s, rinsed twice in 100% acetone for 10 seconds each, rinsed in dH_2O for 10 seconds and baked dry overnight at 42°C. TESPA coated slides were stored dessicated at 4°C with silica gel.

2.4.4. Staining Embryos and Cryostat Sections for β-gal Activity

Cells and embryos were washed in PBS and treated with fix solution (0.2% gluteraldehyde; 5mM EGTA (pH7.3); 2mM $MgCl_2$ in 0.1M sodium phosphate) for 5 minutes (cryostat sections and cells), 15 minutes (8.5d.p.c. embryos) and 30 minutes (9.5-10.5d.p.c. embryos) at room temperature. The samples were then treated with wash buffer (20mM $MgCl_2$; 0.01% deoxycholate; 0.02% nonidet in 0.1M sodium phosphate pH7.3) 3 times for 5 minutes (cryostat sections and cells) or 20 minutes (embryos) at room temperature. Samples were then stained overnight at 37°C using X-gal stain (1mg/ml X-gal; 250mM potassium ferrocyanide; 250mM potassium ferricyanide in wash buffer). The stain was replaced with wash buffer and the samples stored at 4°C.

2.4.5. Haematoxylin and Eosin Counterstaining

Cryostat sections were counterstained after X-gal staining for 2-3 minutes with haematoxylin (stock solution- Sigma, diluted 1/12 in water). The sections were rinsed

clear in water and then treated for 1 minute 0.1% eosin (BDH) in 70% ethanol, and dehydrated by washing 2 times for 5 minutes in 95% alcohol and 2 times 5 minutes in 100% ethanol. The sections were treated for 5 minutes in histoclear, coverslipped and mounted using DPX mountant (BDH) and allowed to dry overnight.

2.4.6. Microscopy and Photography of Specimens

Whole embryos were examined using Olympus SZ40 and Olympus SZH10 dissection microscopes and photographed using an Olympus C-35AD-4 camera.

Sections were examined using an Olympus Vanox AHBT3 microscope and photographed using an Olympus C-35AD-4 camera.

Chapter 3

RESULTS

Characterisation of I114 Reporter Expression Profile and Phenotype

3.1. Introduction

The integration of a gene trap vector into a transcription unit is predicted to produce a fusion transcript consisting of endogenous message upstream of the *lacZ* reporter gene (Figure 1.2B). The consequences of this are threefold: (i) translation of this transcript will produce β–galactosidase (β–gal) activity that correlates with the temporal and spatial expression of the endogenous gene; (ii) the endogenous message will be disrupted downstream of the *lacZ* polyA signal and (iii) the endogenous gene can be identified using 5'RACE-PCR (Skarnes *et al.*, 1992). In this chapter, the β–gal activity profile and, by inference, the expression profile of the endogenous gene trapped in the I114 gene trap integration is described. To identify potential recessive phenotypes associated with gene trap vector insertion, the germline transmission and breeding to homozygosity of the I114 integration is also reported.

3.2. I114 Gene Trap Cell Line

The I114 gene trap cell line was derived from electroporation of the gene trap vector PT1-ATG into the R1 ES cell line (Forrester *et al.*, 1996). Using an *in vitro* pre-screening strategy designed to enrich for genes with developmentally restricted expression patterns, the I114 ES cell line was selected after showing a 2.6 fold induction in β-galactosidase reporter activity after RA treatment for 48 hours. Whole mount *in situ* X-gal staining of tetraploid aggregation chimaeras derived from the I114 ES cells line revealed developmentally restricted reporter activity in the foetal liver

between 9.5-10.5 d.p.c. Southern blot analysis of I114 ES cell DNA digested with EcoRI and hybridised to a *lacZ* probe detects four distinct sized restriction fragments suggesting that four copies of PT1-ATG have inserted into the I114 cell line (Forrester *et al.*, 1996).

C57/Bl6 blastocysts were injected with the I114 ES cell line and transferred to the uteri of pseudopregnant mothers by Jan Ure (CGR). Using coat colour to assess ES cell contribution to offspring, four chimaeric males (48, 49, 50 and 51) were identified which were subsequently testcrossed onto MF1 females. All four chimaeras transmitted the gene trap integration through the germline as judged by coat colour and tail DNA genotyping of testcross offspring. Animals heterozygous for the I114 integration were used for subsequent backcross matings and for the generation of heterozygous and homozygous animals for detailed reporter gene expression.

3.3. I114 Expression Analysis

3.3.1. Embryonic Expression

Embryos heterozygous for the I114 gene trap intergration displayed β–gal activity in the developing liver between 9.5-10.5 d.p.c., agreeing with the original data from aggregation chimaeras (Forrester *et al.*, 1996). To examine in more detail the extent of β–gal activity throughout embryogenesis, X-gal staining of whole mount and cryostat sectioned embryos from 8.0 d.p.c. through to 17.5 d.p.c. was performed. The data shown is generated from embryos heterozygous for the I114 gene trap integration after five generations of backcrossing to the C57BL/6 background unless otherwise stated. The reported β-gal activity profile is also observed in I114 homozygous embryos from (129xMF1)F1 intercrosses at all stages and I114 heterozygous embryos from (129x129)F3 intercrosses between 9.5 and 12.5 d.p.c. (data not shown). Wild type siblings from intercross and backcross litters were used to

control for background β–gal activity. No background activity was observed in control embryos.

The earliest detectable β–gal activity was observed in the 9 somite stage embryo (8.0-8.5 d.p.c.) in the region destined to form the liver diverticulum immediately posterior to the heart (Figure 3.1A). The ventral view of the same embryo resolves the β–gal activity as two distinct lateral strips in the ventral endoderm of the foregut pocket prior to the fusion of this ventral region to form the foregut tube (Figure 3.1B). At the 12 somite stage (8.5 d.p.c.), as closure of the ventral surfaces of the foregut proceeds caudally, the lateral strips of β–gal activity merge at the anterior end (corresponding to the posterior most extent of foregut closure), resulting in a 'horseshoe' of β–gal activity (Figure 3.1C). Following closure of this region of the foregut, β–gal activity is maintained in the liver diverticulum and is seen at high levels at 10 d.p.c. (27-29 somites) in the foetal liver (Figure 3.1D). From 10 d.p.c., β–gal activity in the developing liver is maintained throughout gestation and in neonates (data not shown).

Cryostat sections of 11.5 d.p.c. foetal livers show reporter activity restricted to hepatic parenchyme and not in the haematopoetic cells which have colonised the foetal liver by this stage (Figure 3.2A). The haematopoetic cells can be identified both morphologically and by their location in the liver sinusoids (Figure 3.2A). This data in conjunction with the fact that I114 reporter activity is observed in the foetal liver prior to its colonisation by haematopoetic cells suggests that I114 reporter activity is restricted to the hepatic lineage.

As gestation proceeds the first additional site of β–gal activity outwith the liver is observed in the region of the follicles of vibrissae primordia in the upper lip at 15.5 d.p.c. (Figure 3.2B). The follicles of vibrissae will give rise to the hair follicles of the whiskers. By 17.5 d.p.c., activity is still observed in this region (data not shown) with additional activity identified in the dorsal root ganglia of the peripheral nervous system (Figure 3.2C). The dorsal root ganglia represent discrete aggregations of primary sensory neurone cell bodies.

Figure 3.1: I114 reporter activity during early liver development

Whole mount X-gal staining of I114 heterozygous mouse embryos

A. Lateral view of 9 somite stage embryo (8.0-8.5d.p.c.); the arrow highlights the β-gal activity in the region immediately posterior to the developing heart. Scale bar 300μm.

B. Ventral view of 9 somite stage embryo (same embryo as A.); β-gal activity resolves as two lateral strips in the foregut pocket (arrowed). Scale bar 300μm.

C. Ventral view of 12 somite stage embryo (8.5dpc); as foregut closure proceeds caudally (black arrow) β-gal activity fuses at the posterior most extent of foregut closure (white arrow). Scale Bar 300μm.

D. Lateral view of 10d.p.c. embryo (28 somites); β-gal activity restricted to the foetal liver (arrowed). Scale bar 650μm

h, developing heart; ys, yolk sac.

Figure 3.1

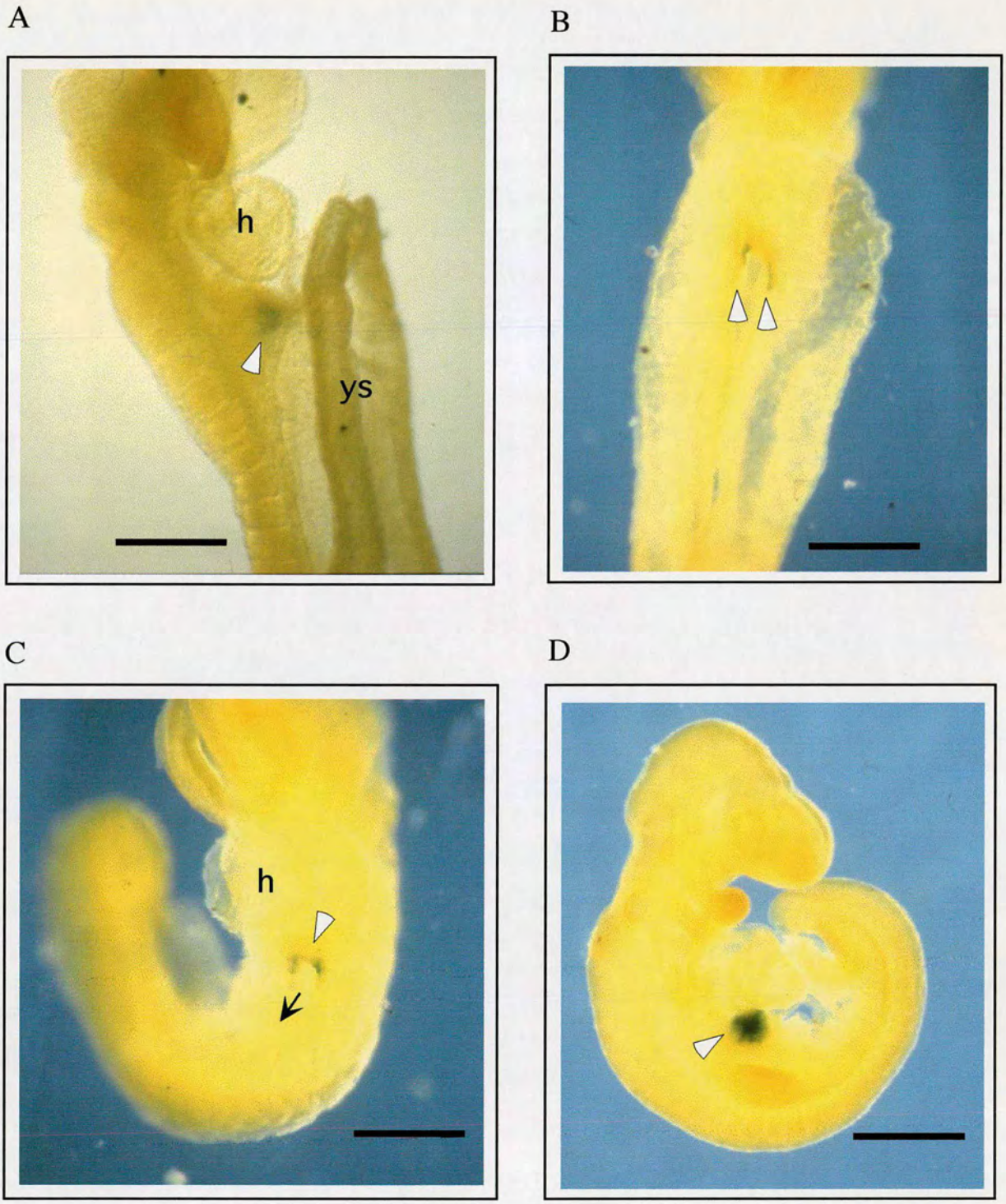
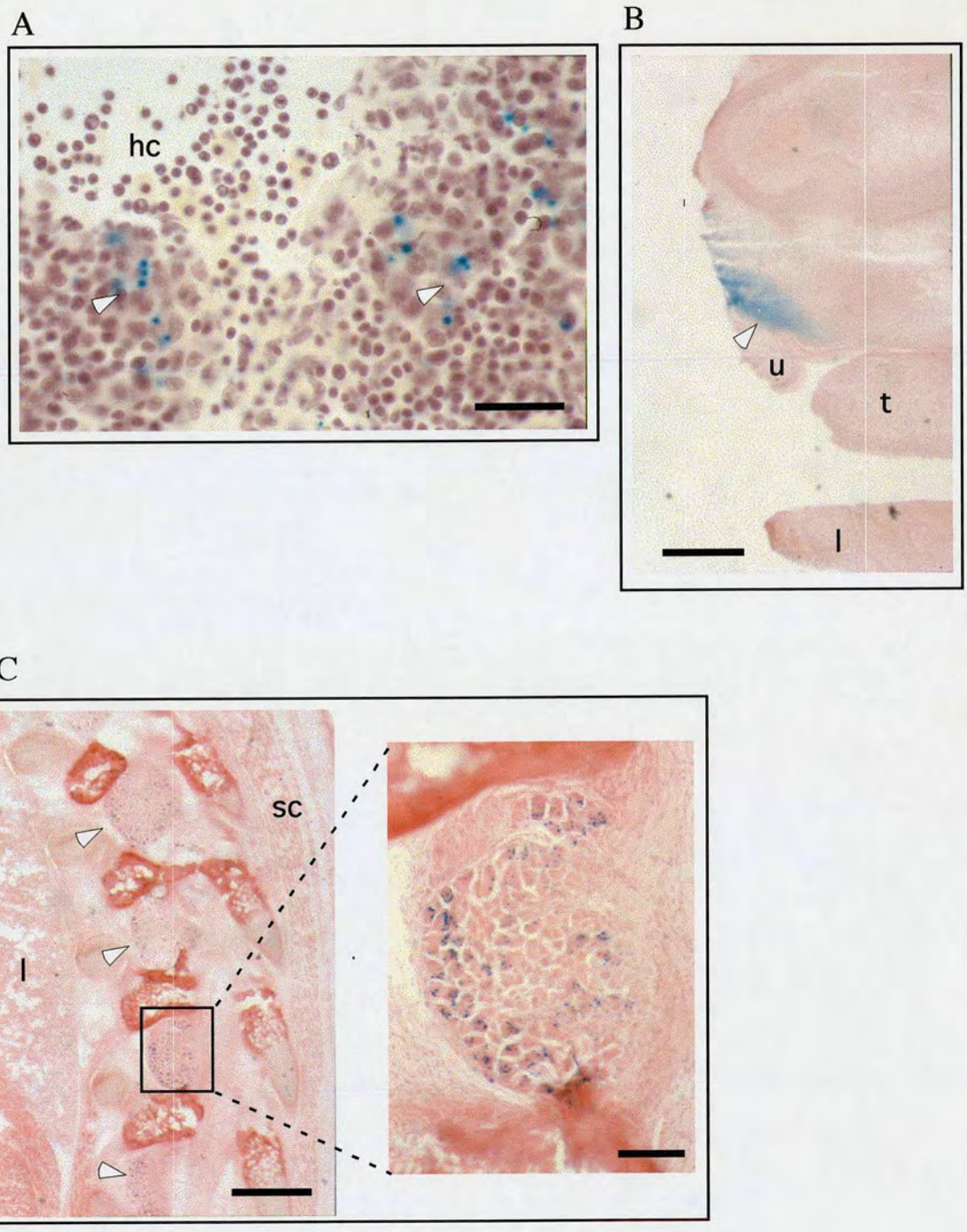


Figure 3.2: I114 reporter activity from mid-late gestation

X-gal staining of croystat sections from I114 heterozygous mouse embryos:
A. Transverse section of 11.5 d.p.c. liver showing reporter activity restricted to the parenchymal cells (arrowed) and not the haematopoeitic cells. Scale Bar 30μm.
hc, haematopoeitic cells.
B. Sagittal section of 15.5 d.p.c. head. Reporter activity can be seen in the upper lip in the region of the follicles of vibrissae (arrowed). Scale Bar 500μm.
u, upper lip; t, tongue; l, lower lip.
C. Sagittal sections of 17.5 d.p.c. embryo. Reporter activity is localised to the dorsal root ganglia (arrowed). Scale Bar 500μM. Higher magnification of boxed area shows in more detail β-gal activity in the dorsal root ganglia. Scale Bar 100μm.
l, lung; s, spinal cord.

Figure 3.2



3.3.2. Adult Expression

To examine the extent of β-gal activity in adults, X-gal staining was performed on cryostat sections from a range of tissues from 3-5 month old I114 homozygous animals on the outbred MF1 background. The same tissues from age matched wild type animals were processed in parallel to control for background β-gal activity. No background β-gal activity was observed in any of the tissues examined.

Sectioning of the adult liver revealed reporter activity in isolated single cells. The exact identity of these cells is difficult to ascertain from the poor cellular morphology presented after cryostat sectioning (Figure 3.3A). The ovaries showed reporter activity which is restricted to the oocytes of the developing follicles (Figure 3.3B). Although Figure 3.3B shows reporter activity only in the oocyte of a secondary follicle, activity has been observed in the oocyte throughout follicle maturation (data not shown). Transverse sections of the testes shows low level, punctate staining of β-gal activity in the germinal epithelium of the seminiferous tubules (Figure 3.3C). In the seminiferous tubules of the testes, spermatogenesis occurs radially from the basal membrane towards the lumen where mature sperm is released. Sertoli cells support this process and can be found throughout the germinal epithelium, although the nucleii are localised to the basal membrane. I114 reporter activity is seen throughout the radius of the germinal epithelium but it is difficult to assign to a specific cell type. Double staining with X-gal and markers for the cell types of the seminiferous tubules and stages of spermatogenesis would help to resolve this question. Examination of neonate ovaries and testes failed to detect any reporter activity. Longitudinal sections of the kidney displays a significant level of reporter activity restricted to the pelvic region as it enters the ureter (Figure 3.3D). Little is known of any distinct function of this tissue other than to collect urine. Sectioning and staining of the heart, gut, lung, spleen and skeletal muscle failed to detect reporter activity in these tissues (data not shown).

Figure 3.3: I114 adult reporter activty

X-gal staining of cryostat sections of adult I114 homozygous tissues.

A. Sectioning of the liver reveals reporter activity restricted to single, isolated cells of hepatic parenchyme (arrowed). Scale Bars 500μm and 50μm.

B. Sectioning of the ovary with reporter activity in secondary follicle. Scale Bar 200μm. Higher magnification of boxed area shows reporter activity in the oocyte of the secondary follicle. Scale Bar 50μm.

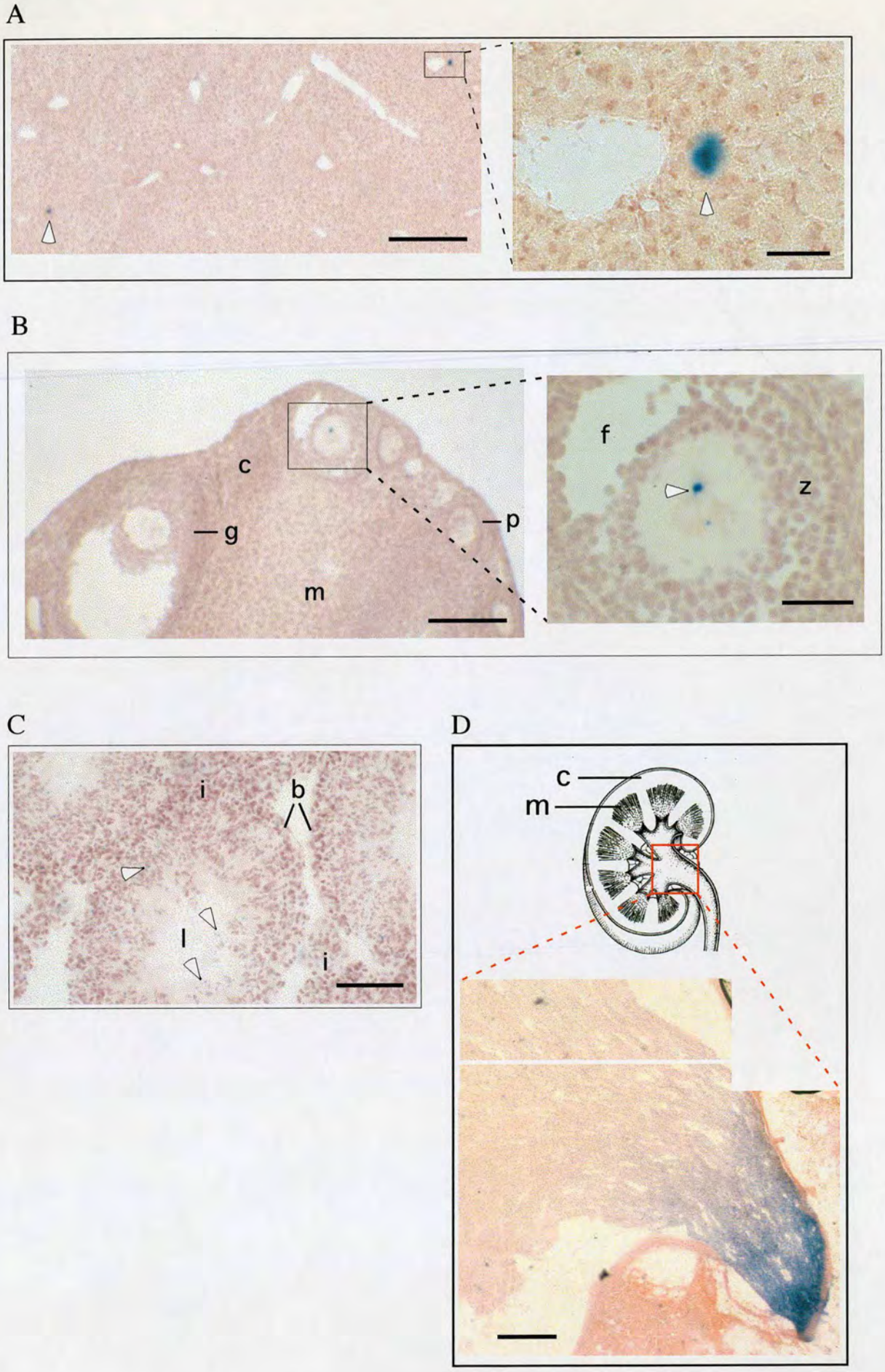
c, cortex; m, medulla; g, Graafian follicle; p, primordial follicle; f, follicular antrum; z, zona granulosa.

C. Longitudinal section of testes with low level staining (arrowed) throughout the seminiferous tubules. Scale Bar 100μm.

i, interstitial cells; l, lumen; b, basal membrane.

D. Longitudinal section of the kidney with staining in the collecting ducts of the pelvic region. Scale Bar 500μm.

Figure 3.3



3.4. Comparison of I114 Reporter Activity With AFP Expression

3.4.1. Embryonic Expression

I114 embryos display β–gal activity in the ventral endoderm of the foregut pocket prior to any morphological signs of liver development. To correlate this specific activity to a well characterised marker of early hepatic differentiation, I114 β–gal activity was compared to expression of AFP (Shiojiri, 1981; Gualdi *et al.*, 1996).

Animals carrying a targeted allele of AFP were generated by homologous recombination in ES cells. A β-geo cassette was inserted between exon 1 and 3, disrupting AFP function and placing β-geo under the control of the AFP promoter (Philippe Gabant, manuscript in preparation). β–gal activity in heterozygous animals ($AFP^{(\beta\text{-geo}/+)}$) will mimic the expression of AFP, thus providing a simple, comparable assay of AFP expression relative to I114 β–gal activity. All comparisons of I114 heterozyogus and $AFP^{(\beta\text{-geo}/+)}$ were carried out with wild type embryos from both litters used as controls.

Whole mount *in situ* X-gal staining of 6 somite stage embryos shows no I114 activity. At this stage, AFP expression is observed as patchy expression in the visceral endoderm (Figure 3.4A). At the 9 somite stage (8.0-8.5 d.p.c.), when I114 (+/-) β–gal activity is first detected in the foregut pocket of the embryo, $AFP^{(\beta\text{-geo}/+)}$ reporter activity is still limited to extra-embryonic tissue (Figure 3.4B). AFP reporter activity is first observed in embryonic tissue at 8.5 d.p.c. (15 somites) in the liver diverticulum and is also maintained in the yolk sac (Figure 3.4C). By 11.5 d.p.c., $AFP^{(\beta\text{-geo}/+)}$ embryos show high levels of reporter activity in the liver and yolk sac (Figure 3.4D). No other embryonic stages were examined for reporter activity.

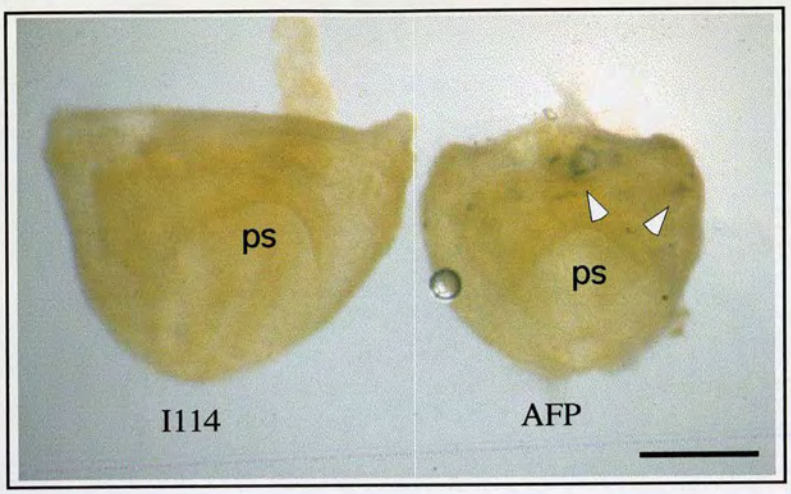
In all embryonic stages examined, $AFP^{(\beta\text{-geo}/+)}$ embryos show patchy reporter activity in the yolk sac. However, whole mount *in situ* hybridisation of wild type embryos with an antisense AFP riboprobe shows ubiquitous expression throughout the yolk sac (Philippe Gabant, manuscript in preparation). The discrepancy between

Figure 3.4: Comparison of I114 and AFP reporter activity during embryogenesis.

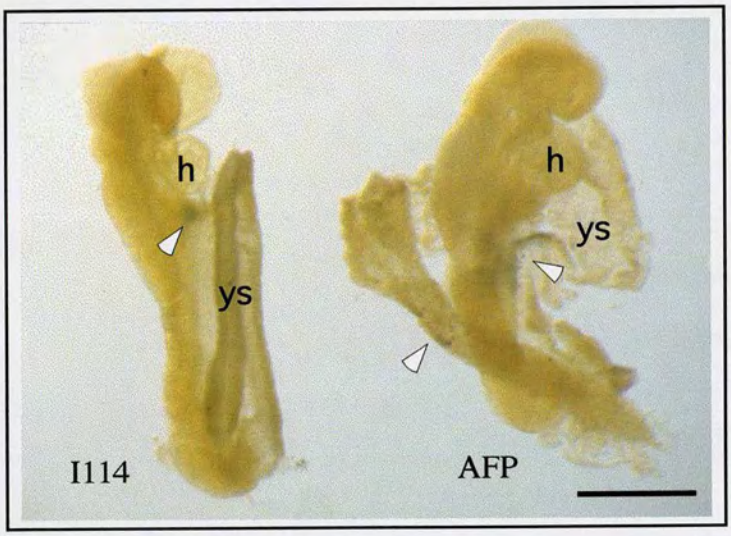
X-gal staining of whole mount I114 heterozygous and AFP $^{(\beta\text{-geo}/+)}$ embryos.
A. Posterior view of 6 somite embryos (8.0 d.p.c.). No staining is observed in the I114 heterozygotes. AFP $^{(\beta\text{-geo}/+)}$ embryos show staining in the yolk sac (arrowed). Scale Bar 330μm.
ps, primitive streak.
B. Lateral view of 9 somite embryos (8-8.5 d.p.c.). As reported previously, β-gal activity is seen in the presumptive hepatic endoderm in I114 heterozygotes. Same stage AFP $^{(\beta\text{-geo}/+)}$ embryos have activity solely in the yolk sac. Scale Bar 500μm.
h, heart; ys, yolk sac.
C. Lateral view of 15 somite AFP $^{(\beta\text{-geo}/+)}$ embryo showing reporter activity in the liver diverticulum and yolk sac (arrowed). Scale Bar 660μm.
D. Lateral view of 11.5 d.p.c. AFP $^{(\beta\text{-geo}/+)}$ embryo with reporter activity in the foetal liver and yolk sac (arrowed). Scale Bar 1.33mm.

Figure 3.4

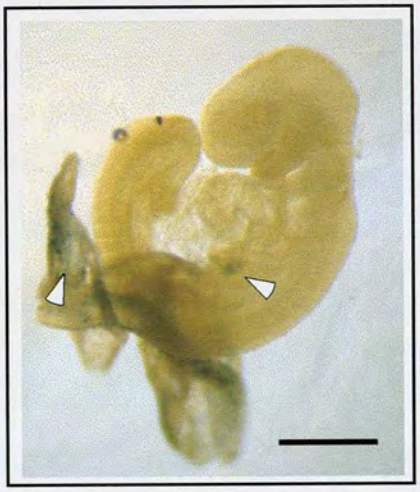
A



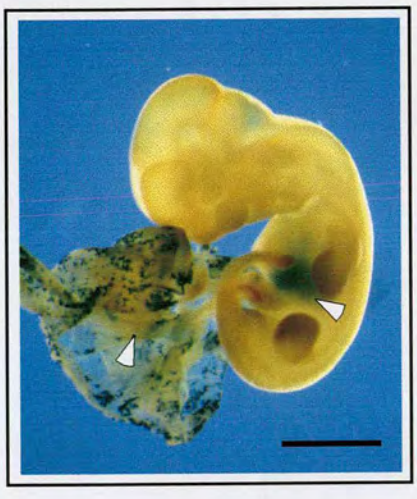
B



C



D



AFP$^{(\beta\text{-geo}/+)}$ activity and wild type AFP expression is possibilty a result of the disruption of a yolk sac specific promoter in the first intron of the AFP gene by the targeting vector (Philippe Gabant, manuscript in preparation). A good correlation is seen between AFP$^{(\beta\text{-geo}/+)}$ reporter activity and wild type AFP expression in the foetal liver (Philippe Gabant, manuscript in preparation).

In conclusion, comparing I114 and AFP$^{(\beta\text{-geo}/+)}$ reporter activity has shown that I114 reporter activity is detected earlier in the foetal liver than AFP and that this expression is specific to embryonic tissue with no significant I114 β–gal activity detected in the yolk sac.

3.4.2. Adult Expression

A comparison of I114 heterozygous and AFP$^{(\beta\text{-geo}/+)}$ reporter activity was carried out on age matched tissues from 3 month old mice to identify common regions of expression in the adult. Wild type littermates of both I114 (+/-) and AFP$^{(\beta\text{-geo}/+)}$ animals were used to control for background β-gal activity. The reporter activity of I114 heterozygous animals matched that of the I114 homozygous animals as described previously (data not shown). In common with I114 heterozygotes, AFP$^{(\beta\text{-geo}/+)}$ reporter activity is observed in the adult liver in isolated regions of the hepatic parenchyme. However, activity is more widespread than the single cells observed for I114 heterozygotes (Figure 3.5A compared to Figure 3.3A). The only other common site of expression in the adult was in the testes. Longitudinal sections of AFP$^{(\beta\text{-geo}/+)}$ testes reveals punctate reporter activity in the seminiferous tubules which, in comparison to I114 heterozygotes (Figure 3.3C), appears more widespread and at a higher level (Figure 3.5C). Expression was also seen, restricted to single cells, of the crypts of Lieberkuhn in the small intestine (Figure 3.5B). No reporter activity is observed in the small intestine of I114 heterozygotes (data not shown). Background β-gal activity was observed in AFP$^{(\beta\text{-geo}/+)}$, I114 heterozygous and wild type animals in the cortex and medulla of the ovary and the cortex and medulla of the kidney (data not

Figure 3.5: AFP $^{(\beta\text{-geo}/+)}$ adult reporter activity

X-gal staining of cryostat sections from AFP $^{(\beta\text{-geo}/+)}$ adult tissues.

A. Reporter activity is observed in the hepatic parenchyme of AFP $^{(\beta\text{-geo}/+)}$ livers. Scale Bar 100μm.

B. Longitudinal section of AFP $^{(\beta\text{-geo}/+)}$ small intestine reveals reporter activity in a single cell of the crypts of Lieberkuhn. Scale Bar 100μm.

v, villi; m, muscularis mucosae; c, crypts of Lieberkuhn.

C. Longitudinal section of AFP $^{(\beta\text{-geo}/+)}$ testes with reporter activity in the seminiferous tubules. Scale Bar 100μm.

i, interstitial cells; l, lumen of seminiferous tubules; b, basal membrane.

Figure 3.5

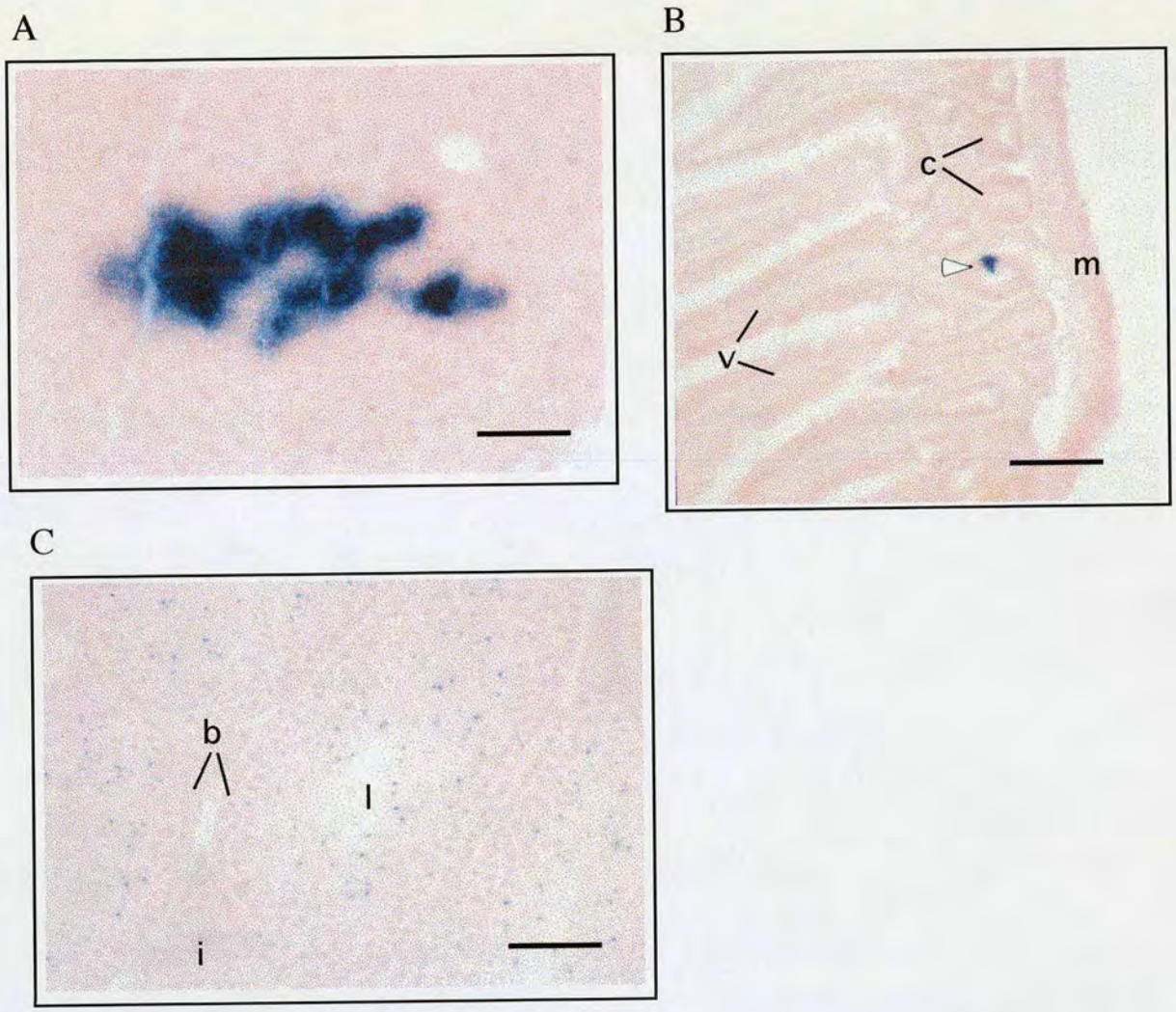


Table 3.1: Comparison of I114(+/-) and AFP$^{(\beta geo/+)}$ β-gal activity

Embryonic	β-gal activity	
Stage/ tissue	I114(+/-)	AFP$^{(\beta geo/+)}$
6.0 d.p.c.		
v.endoderm	-	+
8.5 d.p.c.		
v.endoderm	-	+
def. endoderm	+	-
9.5-11.5 d.p.c.		
v.endoderm	-	+
liver	+	+

Adult	β-gal activity	
Tissue	I114(+/-)	AFP$^{(\beta geo/+)}$
gut	-	+
heart	-	-
kidney	+*	- *
liver	+	+
lung	-	-
ovary	+*	-*
sk' muscle	-	-
spleen	-	-
testes	+	+

v.endoderm, visceral endoderm; def. endoderm, definitive endoderm; sk' muscle, skeletal muscle.
* high level of background activity in these tissues.

shown). However, no overlap was seen between the I114 specific activity in the oocytes and the pelvic region of the kidney and this background activity. The presence of background β-gal activity in I114 heterozygous, $AFP^{(\beta\text{-geo}/+)}$ and wild type ovaries and kidneys is at odds with the results from the sectioning of I114 homozygous and wild type tissues which showed no such background activity (Section 3.3.2.). This probably reflects unmonitored variations in the pH of the different solutions used for the X-gal staining protocols in Section 3.3.2 and Section 3.4.2. If the solutions used in Section 3.4.2. had a lower pH, this would increase the likelihood of background mammalian β-gal activity which has a more acidic pH optimum compared to the higher optimal pH for bacterial β-gal activity (Alam and Cook, 1990).

Sectioning and staining of spleen, heart, skeletal muscle and lung failed to detect reporter activity. A summary of the comparison between the reporter activity of I114 heterozygotes and $AFP^{(\beta\text{-geo}/+)}$ animals is given in Table 3.1

3.5. I114 Reporter Activity During Tumourigenesis

A limited study was undertaken to assess if, like other markers expressed in the foetal liver, (e.g. AFP; Abelev *et al.*, 1971; Chen *et al.*, 1997), I114 reporter activity is upregulated during liver tumourigenesis.

Diethylnitrosamine (DEN) is a carcinogen of the liver predominantly inducing hepatic adenomas and hepatocellular carcinomas (Moore *et al.*,1981; Drinkwater and Ginsler 1986). DEN was injected by Chris Kemp (Beatson Institute, Glasgow) interperitoneally into 12 I114 homozygous animals at 3 weeks of age. Total livers were dissected from each animal and tumours identified as abnormal foci of cells on the surface of the liver (Drinkwater and Ginsler, 1981). Livers were subsequently frozen for cryostat sectioning and X-gal staining. Ten months after DEN injection, 3 out of the 5 animals sacrificed showed gross signs of tumourigenesis in the liver. On sectioning the 3 tumour-bearing livers, 2 displayed an modest increase (Figure 3.6A) and the third a massive increase in β-gal activity (Figure 3.6B) within undefined

Figure 3.6: I114 reporter activity during tumourigenesis

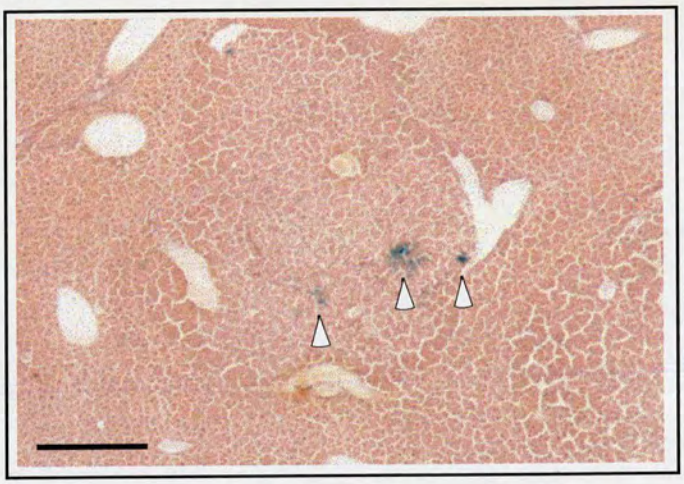
X-gal staining of cryostat sections from two individual I114 homozygous animals (10 months after DEN injection)

A.+B. Sectioning of liver from two separate animals shows a modest increase in reporter activity (arrowed) in comparison to normal liver (Figure 3.3A). Scale Bar A=200μm, B=100μm.

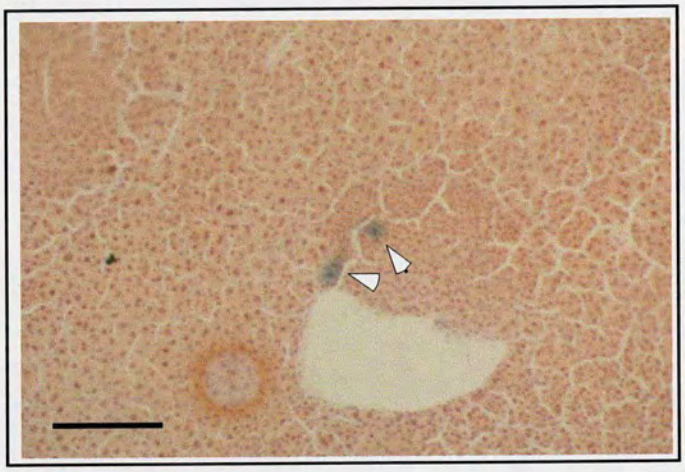
C. Sectioning of liver shows a massive increase in reporter activity over normal liver (Figure 3.3A). Scale Bar 500μm.

Figure 3.6

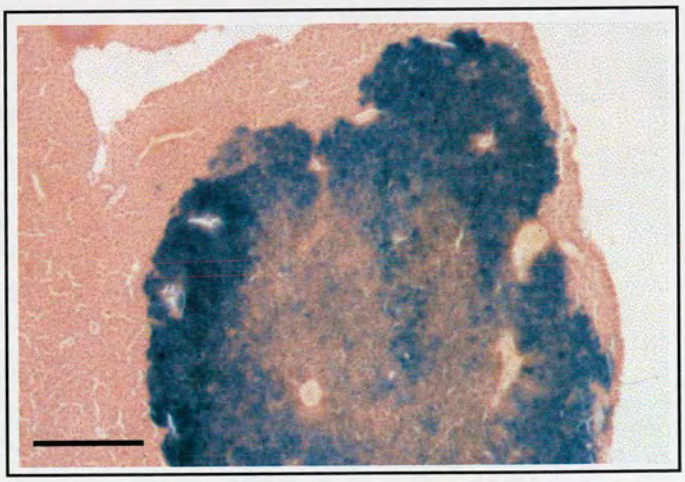
A



B



C



regions of the hepatic parenchyme when compared to an untreated adult (Figure 3.3A). This apparent induction in β-gal activity could be an artefact of the mutagenesis. However, DEN induces DNA point mutations (Drinkwater and Ginsler, 1986) so it would perhaps be unlikely that the induction in β-gal activity is due to the clonal expansion of a mutagenic event which, for example, causes the ubiquitous expression of the reporter gene.

Obviously, little can be concluded from such a limited study. To confirm that the induction in β-gal activity marks hepatic tumours, the study would need to be repeated on a larger scale with suitable controls. Wild type animals injected with DEN and untreated I114 and wild type animals would be processed in parallel. A comparison of the I114 β-gal activity with the expression of the endogenous trapped gene (see Chapter 6) and hepatic tumour markers would also help to define nature of the β-gal activity.

3.6. I114 Breeding Analysis

Four chimaeric male offspring derived from the injection of I114 ES cells into blastocyst stage embryos transmitted the gene trap integration into the germline. Chimaeras were subsequently backcrossed onto different genetic backgrounds to examine the viability of animals homozygous for the I114 integration. Chimaeras 48 and 51 were backcrossed onto the MF1 outbred genetic background and chimaera 49 was backcrossed onto the inbred 129/CGR and C57BL/6 backgrounds. The majority of the genotyping work was carried out by Tony Coyle and Dianne Peddie.

On the MF1 outbred background, (129xMF1)N1 generation heterozygote offspring were intercrossed. Genotyping of intercross litters was performed by Southern blotting of tail genomic DNA restriction digested with BglII followed by hybridisation to the 500bp *en-2* fragment of the gene trap vector. This resolves a fragment of approximately 3kb corresponding to the endogenous genomic copy of *en-2* and 4 distinct fragments greater than 10kb corresponding to the integrated copies of the

gene trap vector (Figure 3.7). The endogenous *en-2* fragment provides a loading control to aid the genotyping of I114 heterozygous and homozygous animals based on the signal intensity of the *en-2* hybridising vector fragments (Figure 3.7). Animals not readily genotyped by this method were testcrossed onto wild type animals to distinguish between heterozygotes and homozygotes. Table 3.2 shows that animals homozygous for the I114 gene trap integration are produced at the expected Mendelian ratio of 1:2:1 (wild type: heterozygous: homozygous) at weaning stage. Moreover, intercrossing of I114 homozygous animals has produced a homozygous line.

Intercrosses were set up after 5 generations of backcrossing the gene trap integration onto the C57BL/6 inbred genetic background. Reliable genotyping of these intercrosses was carried out using the dot blot technique which was developed in the laboratory by Dianne Peddie. (Section 2.1.4.4.). As with intercrosses for the other backgrounds, animals homozygous for the gene trap integration showed no overt phenotype being produced at the expected frequency (Table 3.2). A viable homozygous line has been set up on the C57BL6 background by intercrossing F5 generation I114 homozygotes. To date 11 litters have been born with an average of 4.5 pups per litter, which is within the expected litter size for the inbred genetic backgrounds.

Intercrosses on the 129/CGR background were set up after backcrossing to the N3 and N4 generations. To quickly assess the genotype ratio of the 129/CGR intercross offspring, the numbers of I114 heterozygous and homozygous offspring were pooled and compared to the numbers of wild type offspring. At weaning, 52 I114 heterozygous and homozygous versus 19 wild type offspring were genotyped which is in keeping with the expected Mendelian ratio of 3:1, wt: heterozygote and homozygote.

Therefore, the integration of the gene trap vector in the I114 cell line has no overt effect on viability or fertility when bred to homozygosity on the outbred MF1 and inbred 129/CGR and C57BL/6 genetic backgrounds.

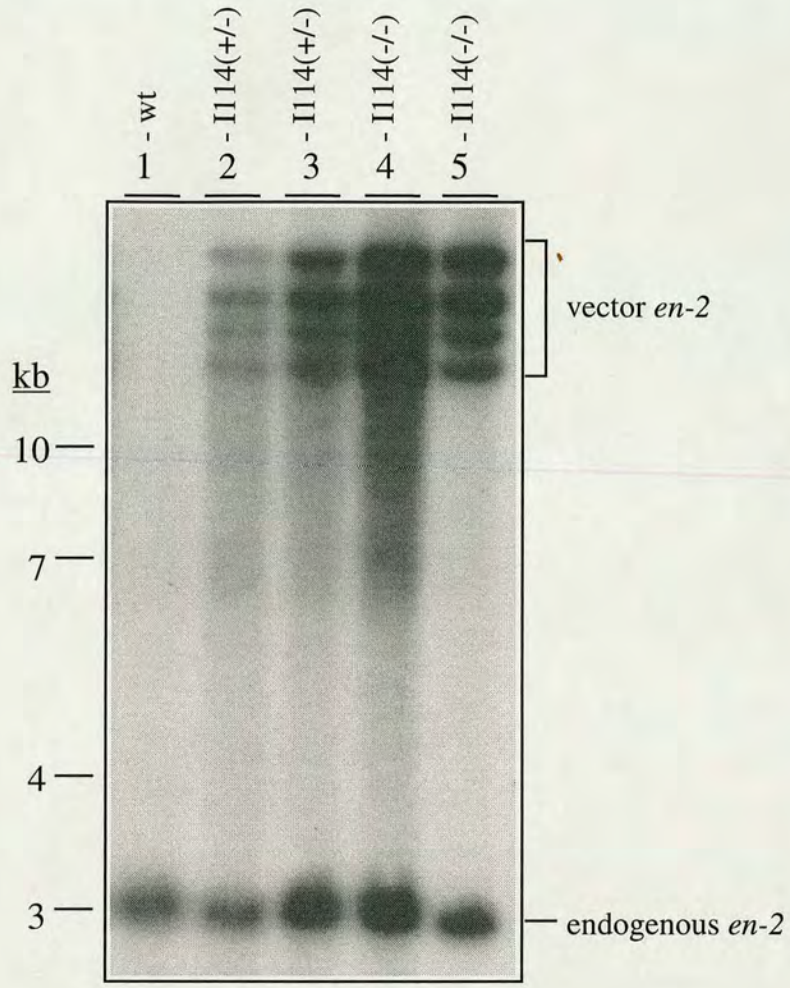


Figure 3.7: Southern blot analysis of I114 genomic DNA.

Genomic DNA from I114 heterozygous intercross litters digested with BglII and hybridised to en-2 exon probe. The four positively hybridising bands greater than 10kb in lanes 2-5 correspond to gene trap vector copies. The band at 3kb corresponds to endogenous en-2 and serves as a DNA loading control. For example, comparison of lanes 3 and 4 shows equal DNA loading. Lane 4 has a greater signal intensity of the vector copies indicative of an I114 homozygous animal.

Table 3.2: Genotype of I114 heterozygous intercross litters at weaning

Genetic Background	wild type	I114 heterozygotes	I114 homozygotes
(129 x MF1) F1 I-X	33	79	32
(C57Bl/6 x C57Bl6) F5 I-X	10	22	12

3.7. Discussion

Does I114 Define Hepatic Specification?

The I114 β–gal activity observed in the ventral foregut endoderm corresponds to the location of the presumptive hepatic endoderm identified in previous studies (Houssaint, 1980; Gualdi *et al.*, 1996). In the explant studies of Houssaint (1980), only this presumptive hepatic endoderm isolated from the 9 somite mouse embryo onwards was capable of differentiation into morphologically distinct hepatocytes on culturing with chick hepatic mesenchyme. Accordingly, this study suggests that the hepatic lineage has been specified, at earliest, by the 9 somite stage. In another study by Gualdi *et al.* (1996), hepatic specification was defined by expression of serum albumin using an RT-PCR assay and was determined to have occurred by the 7-8 somite stage. This slightly earlier timing of hepatic specification compared to Houssaint (1980) may reflect the different assays used to define specification in this study.

The earliest β–gal activity observed in I114 embryos, at the 9 somite stage, corresponds well to the perceived threshold of hepatic specification observed in the study by Houssaint (1980). Moreover, it is likely that I114 reporter gene transcription is activated earlier than at the 9 somite stage when protein activity is detected. Consequently, this would correlate well to the specification of the presumptive hepatic endoderm as determined by the transcriptional activation of liver specific genes as seen in the study of Gualdi *et al.* (1996).

The reporter activity associated with the I114 integration is observed prior to the thickening of the foregut endoderm to form the liver diverticulum, the first morphological signs of liver differentiation. Moreover, it suggests that hepatic specification occurs before the ventral closure of the foregut endoderm in two apparently distinct cell populations. This has not previously been reported. Specification of the hepatic lineage from gut endoderm is dependent on inductive signals from the adjacent cardiac mesoderm from the 4 somite stage (Le Douarin, 1975;

Gualdi *et al.*,1996). At this stage, and up to the 9 somite stage, the foregut endoderm has yet to fuse ventrally with two populations of cells existing adjacent to inductive cardiac mesoderm. I114 reporter activity nicely illustrates that hepatic specification of the gut endoderm occurs at such an early developmental stage due to its position relative to cardiac mesoderm, rather than from a intrinsic property of an endodermal cell population.

The comparison of the earliest detectable I114 and AFP$^{(\beta\text{-geo}/+)}$ reporter activities suggests that the I114 gene trap integration provides the earliest, most specific marker of hepatic specification and development identified to date. Subsequently, it is appealing to propose that the trapped gene associated with this activity is one of the first genes to be activated as a result of hepatic specification, which to some degree defines the hepatic phenotype. Consequently, the I114 gene trap cell line would provide a simple and cheap assay of hepatic specification which could be used in tissue explant culture studies or in the phenotype analysis of future gene knock-outs affecting hepatic specification.

Previous studies have highlighted that, as with I114 reporter activity, expression of both AFP and *alb* in the definitive endoderm occurs prior to morphological signs of liver development (Gualdi *et al.*, 1996; Shiojiri, 1981; Cascio and Zaret, 1991). However, unlike both AFP and *alb*, which are expressed in the visceral and definitive endoderm during liver development (Meehan *et al.*, 1994; Dziadek and Adamson, 1978), I114 reporter activity is restricted to the definitive endoderm. Given that visceral endoderm is continuous with the emerging definitive endoderm during hepatic specification, markers such as AFP and *alb*cannot distinguish between the two lineages at this stage. I114 β-gal activity, being exclusive to the definitive endoderm, affords excellent definition of the morphology of liver specification and development from the definitive endoderm at this early stage.

In the study of reporter activity in AFP$^{(\beta\text{-geo}/+)}$ embryos, activity is restricted to the liver diverticulum from the 15 somite stage onwards. In the study of Gualdi *et al.* (1996), AFP expression is detected by RT-PCR throughout the anterior and posterior

definitive endoderm from the 4 somite stage which was subsequently upregulated in the liver diverticulum and down regulated in the rest of the gut endoderm. The difference observed in AFP expression between these two studies may reflect the higher sensitivity of the RT-PCR assay compared to X-gal staining for β-gal activity. However, there are two other possibilities that could explain the discrepancy between the two studies. Firstly, in the study by Gualdi *et al.* (1996), the AFP expression observed in the isolated definitive endoderm could be due to the co-isolation of AFP expressing visceral endoderm, which is continuous with definitive endoderm at this stage. Secondly, the disruption of the AFP allele with the β-geo targeting cassette affects the expression pattern of AFP in the yolk sac by removing a promoter element in the first intron. It is conceivable that this or another promoter element disrupted by the targeting construct is necessary for AFP expression throughout the gut endoderm at this earlier stage.

What Cell Type is Marked by I114 Activity in the Adult?

In the adult, I114 reporter activity is observed in single, isolated cells of the liver. Such a striking pattern of β-gal activity may be a faithful report of endogenous gene expression marking a sub-population of hepatic cells or it may be a consequence of variegated reporter gene expression. In the mouse, pericentromeric and high copy number transgene insertions can contribute to transgene silencing resulting in a mosaic pattern of reporter gene expression in tissues where more widespread expression of the reporter is predicted (Dobie *et al.*, 1997). Although this phenomenon has only been reported in transgenic mice generated by pro-nuclear injection of DNA, it could be causing the highly restricted reporter activity observed in I114 adult livers.

Another alternative is that I114 reporter activity could mark a specific hepatic cell type expressing the endogenous trapped gene. From the profile of I114 reporter activity in the foetal and adult liver, it is intriguing to speculate that the marked cells are the potential stem cells of the adult liver. In normal adult livers the stem cells are

postulated to exist as very small sub-population of phenotypically indistinct cells located in the interlobular bile ducts and canals of Hering. As with other stem cells systems, the actual liver stem cells are defined more by their progeny, the oval cells, than by their own phenotype. Oval cells can be studied during liver regeneration in response to liver damage. Experimentally, a two-thirds partial hepatectomy (2/3 PH) or administration of hepatotoxins such as carbon tetrachloride (CCl_4) induces the proliferation of mature hepatocytes which have a replicative potential approaching that of haematopoietic stem cells (Gerlach *et al.*, 1997; Overturf *et al.*, 1997). However, when this hepatocyte mediated regenerative process is blocked, for example, by administration of 2-acetylaminofluorene (AAF), a proliferation of phenotypic biliary cells known as oval cells is observed around the hepatic portal area. Certain aspects of the oval cell phenotype are similar to foetal liver cells. The proliferating oval cells invade the surrounding hepatic parenchyme where they differentiate into mature hepatocytes. Moreover, oval cell proliferation is associated with an increase in the expression of foetally expressed genes such as AFP and *alb* (Golding *et al.*, 1995). The function of growth factors intrinsic in foetal liver development, for example SF/HGF, are also implicated in the proliferation of oval cells (Alison, 1998).

From this overview, there are several similarities between I114 reporter activity and the characteristics of liver stem cells. Firstly, the highly restricted, infrequent reporter activity observed in I114 adult livers may correlate with the small population of stem cells present in the normal adult liver. Unfortunately, the poor cellular morphology of the cryostat sections makes it difficult to assign I114 reporter activity to individual hepatic cell types or structures. Secondly, there is a foetal phenotype associated with oval cells, the immediate progeny of stem cells. This allows one to postulate that I114 reporter activity, being a foetal marker, may also mark the liver stem cells. By inference, the same would be true of $AFP^{(\beta\text{-geo}/+)}$ reporter activity which is also highly restricted in the adult liver although not to the same degree as I114 activity.

Our preliminary experiment suggesting the induction in I114 reporter activity is associated with hepatocarcinogenesis may also support this idea. Evidence exists for

both hepatocytes and oval cells as the cellular origins of hepatic tumours (Alison, 1998). In different animal models, an increased proliferation of oval cells, as well as the inability of proliferating oval cells to differentiate into hepatocytes (and therefore continue proliferating), is associated with an increased incidence of hepatocellular carcinomas (Isfort *et al.*, 1997; Betto *et al.*, 1996). However, DEN is believed to induce hepatic tumours without oval cell proliferation (Bralet *et al.*, 1996). It is therefore unclear if the induction of I114 reporter activity during tumourigenesis (if indeed there is an induction) is concomitant with an increase in oval cell or hepatocyte proliferation.

Although the above is highly speculative, it would be of considerable interest to identify the nature of the cells with I114 reporter activity. The different morphology and antigenic profile of hepatocytes, bile duct cells and oval cells would allow the specification of I114 reporter activity to a hepatic cell type. For example, using double staining protocols, one could investigate if I114 reporter activity co-localised with the oval cell specific markers SCF, *c-kit* and OV6 antibody (Fujio *et al.*, 1996; Hixson *et al.*, 1997) in liver sections during oval cell proliferation.

Functional Analysis of the Endogenous I114 Gene

Intercrossing of the I114 gene trap integration to homozygosity has no overt effect on the viability or fertility of these animals. Chapters 4 and 6 will show that this lack of phenotype is due to the failure of the I114 gene trap integration to disrupt the function of the endogenous trapped gene. A number of gene trap lines published to date also report the incomplete knock-out of the endogenous gene by the gene trap vector (McClive *et al.*, 1998; Voss *et al.*, 1998; Sam *et al.*, 1998; Faisst and Gruss, 1998). Integrations into *TFEB*, *MAP-4*, *aquarius* and *bodenin* have little or no effect on the wild type expression levels of these genes when the integration is bred to homozygosity. Consequently, these gene trap lines show no phenotype (McClive *et al.*, 1998; Voss *et al.*, 1998b; Sam *et al.*, 1998; Faisst and Gruss, 1998). Expression

of the wild type transcript is predicted to result from the splicing around of the gene trap vector by the endogenous gene, producing sequences 3' of the vector insertion (Voss *et al.*, 1998b; Gasca *et al.*, 1995).

Despite the failure of the I114 gene trap integration to disrupt the endogenous gene, the I114 cell line presents a excellent opportunity to identify and characterise a gene intrinsically involved in liver specification and development. Moreover, I114 reporter activity is an excellent marker of liver ontogeny which could be widely used in other studies of liver development.

Chapter 4

RESULTS

Molecular Characterisation of the I114 Gene Trap Integration

4.1. Introduction

Gene trapping introduces a molecular tag into a locus, facilitating the identification of the endogenous gene it has disrupted (Skarnes *et al* ., 1992). Splicing of the gene trap vector to an endogenous splice donor produces a fusion transcript which allows for the amplification of endogenous gene sequences by 5' RACE-PCR using primers complementary to vector sequences (Frohman *et al.*, 1988; Section 2.1.8.1.). The molecular characterisation of the I114 gene trap integration has proven to be more complex than conventional gene trap integrations. This chapter describes the identification of multiple different fusion transcripts from the I114 line and compares the expression pattern of these fusion transcripts with reporter activity in the I114 embryo.

4.2. I114 Molecular Analysis

Northern blot analysis of I114 ES cell RNA hybridised to a 2.2kb *lacZ* probe (Section 2.1.5.3.) resolves a single fusion transcript of approximately 4kb (Figure 4.1). Furthermore, in agreement with reporter activity, the expression of this fusion transcript is induced in I114 ES cells differentiated in the presence of RA (data not shown, Forrester *et al.*, 1996). The 4kb fusion transcript will comprise 3.4kb of the *en2-lacZ* reporter and the remaining 600bp is likely to be derived from the endogenous trapped gene. This suggests that the vector has inserted towards the 5' end of the endogenous gene as observed for the majority of cell lines isolated in this screen

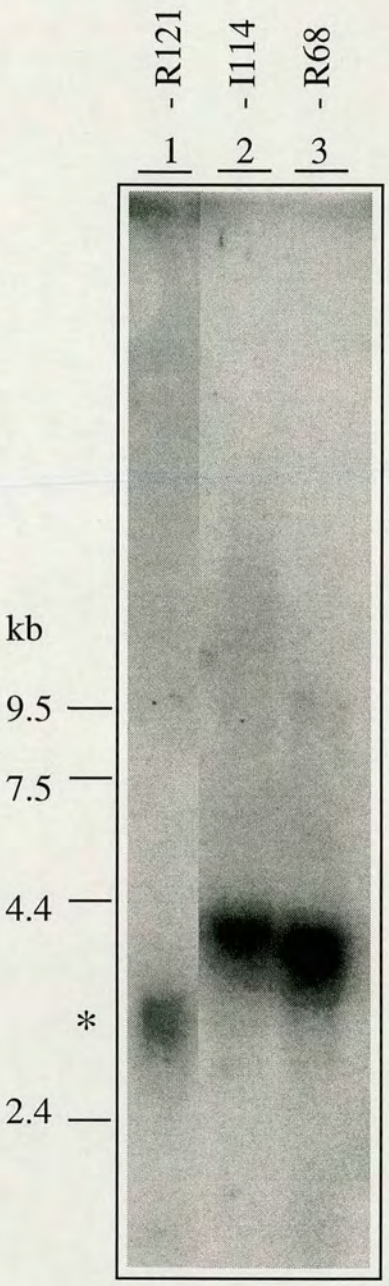


Figure 4.1: Analysis of the I114 fusion transcript by Northern Blot

Detection of fusion transcipts by Northern Blot analysis of total RNA from R121, I114 and R68 gene trap cell lines hybridised to 32^{P} labelled *lacZ* probe.
Cell line R121 (lane 1) expresses the minimum *lacZ* transcript of 3.3kb (asterisk) providing a size standard for comparison to the fusion transcripts of the I114 and R68 cell lines. R68 was included as a known *lacZ*-fusion expressing cell line.
R121, I114 and R68 were all isolated in the RA responsive pre-screen (Forrester et al., 1996).
Markers (GibcoBRL. Cat No. 15620-016).

(Forrester *et al.*, 1995). Hybridisation of I114 ES cell RNA to the *en2* intron probe derived from the gene trap vector (Appendix IV; Section 2.1.5.3.) failed to detect a transcript (data not shown) indicating that the *en-2* splice acceptor is operating efficiently in the I114 gene trap cell line (Townley *et al.*, 1997).

4.2.1. Identification of I114 Fusion Transcripts Using 5'RACE-PCR

(a) ES Cells

5'RACE-PCR (Protocol 2) was performed on RNA isolated from I114 ES cells treated with RA for 48hours to maximise the level of the fusion transcript in the initial RNA sample. Plasmid DNA was isolated from recombinant RACE clones that hybridised to the gene trap vector *en2* probe. Positive clones were rescreened by restriction digest with KpnI and SpeI, and clones containing an insert of greater than 120bp were sequenced. Manual sequencing identified 9 RACE clones containing endogenous sequences correctly spliced to the *en2* splice acceptor site. They were divided into 4 groups (Table 4.1). Six of these clones (ES-α3, α23, β11, β13, β15 and n1:3; Group I) contained between 50 and 70 nucleotides of identical sequence (Figure 4.2) with the remaining three clones (ES-n1:7, n1:2 and α7; Groups II, III, and IV respectively) showing no sequence homology to the Group I clones or each other (Figure 4.2). A further three clones, (ES-β39, β47 and n1:8; Group V) contained *en-2* intronic vector sequence upstream of the splice acceptor, indicative of unspliced transcripts (Table 4.1).

(b) Foetal Liver

Given the presence of multiple, correctly spliced transcripts in I114 ES cell RNA, 5'RACE-PCR was carried out on two different foetal liver RNA samples to analyse the presence and relative abundance of the different fusion transcripts *in vivo*.

TABLE 4.1: Number and size of the different RACE clones isolated from the I114 gene trap line

GROUP	RACE1 (ES cells)		RACE 2 (liver)		RACE 3 (liver)		TOTAL
	No. Clones	Size(bp)	No.Clones	Size(bp)	No.Clones	Size(bp)	
I (L-A8)	6	25-76	2	64+79	13	NA	21 (58%)
II (L-35)	1	54	0	-	3	46-58	4 (11%)
III (ES-n1:2)	1	141	0	-	0	-	1 (3%)
IV (ES-α7)	1	95	0	-	0	-	1 (3%)
V (intron)*	3	>100	2	>100	4	NA	9 (25%)

NA, not applicable, Group I & V clones were identified by hybridisation in this screen and were not sequenced.

* In the RACE 1 and 2 screens, the exact size of the intron containing clones was not determined.

Figure 4.2: I114 fusion transcript sequences

Longest sequence of the fusion transcript Groups I-IV isolated from the RACE-PCR of RNA from I114 ES cells and I114(-/-) 10.5d.p.c. and 13d.p.c. liver RNA. The translation of endogenous fusion transcript sequence is given in all three frames. Translation in-frame with *en-2/lacZ* is highlighted in red. Lower case sequence at the 3' end of each fusion transcript corresponds to the *en-2* exon sequence from the PT1-ATG gene trap vector. (/) corresponds to the splice junction.

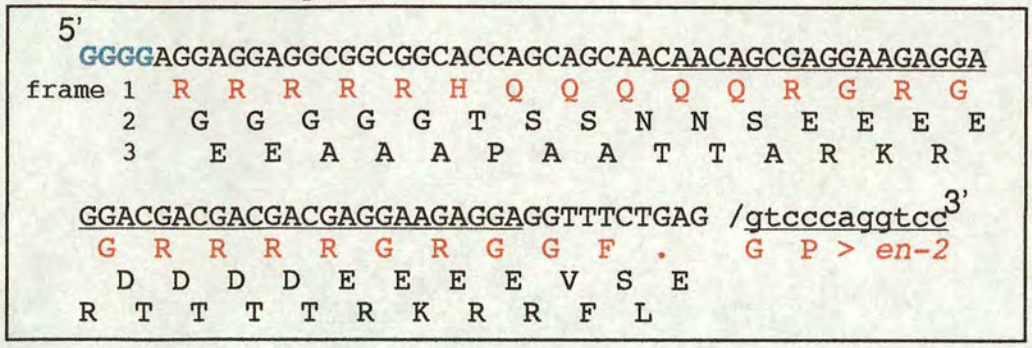
Underlined sequence in L-A8 is complementary to oligonucleotide UBT-1.

G (blue); artificial poly G tail added during RACE-PCR (protocol 2).

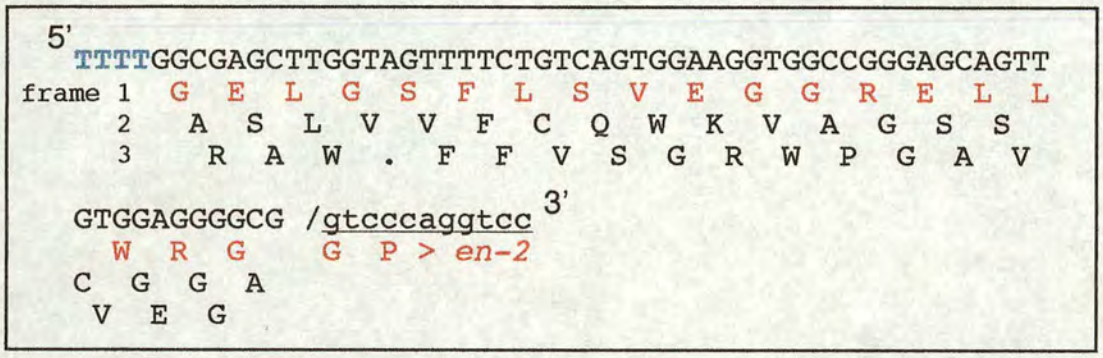
T (blue); artificial poly T tail added during RACE-PCR (protocol 1).

Figure 4.2:

Group I: L-A8 (79bp)

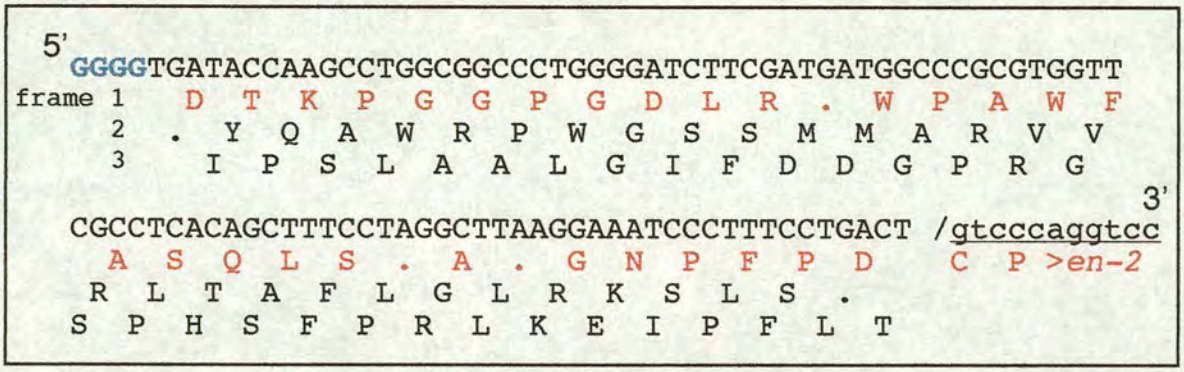


Group II: L-69 (58bp)



Group III: ES-n1:2 (141bp)

Group IV: ES-α7 (95bp)



In the first study, 5'RACE-PCR (Protocol 2) using 10.5d.p.c. I114 homozygous liver RNA isolated only 4 *en-2* positive clones. Clones L-A8 and L-B9 contained sequence identical to Group I clones. Clone L-A8 contained 79bp of novel sequence spliced to the *en-2* exon making it the largest Group I clone isolated to date (Figure 4.1). The two remaining RACE clones, L-A20 and L-A46, contained *en-2* intronic sequence (Group V, Table 4.1).

To isolate a larger sample of RACE clones from the foetal liver, 5'RACE-PCR (Protocol 1) was repeated on liver RNA from 13d.p.c. I114 homozygotes. The 2nd round of PCR was carried out using Pfu polymerase (Stratagene) to produce blunt ended PCR products for cloning into the pCR-Blunt™ vector using the for ZeroBlunt™ PCR Cloning Kit (Invitrogen). A total of 50 *en-2* positive clones were identified by colony PCR. The clones were characterised either by hybridisation to an oligonucleotide probe (UBT-1; Section 2.1.5.3.) complementary to L-A8, the most abundant sequence (Group I; Figure 4.2), or by sequencing. Of the clones large enough to contain an endogenous sequence upstream of the splice acceptor, 13 clones hybridised to the UBT-1 probe (Group I), 4 clones corresponded to unspliced vector sequence (Group V) and 3 clones, L-30, L-35, L-69, identified sequence comparable to clone ES-n1:7 (Group II; Figure 4.2; Table 4.1). The largest of these 3 clones, L-69, contained 58bp of endogenous sequence (Figure 4.2).

(c) Direct Sequencing

Direct sequencing was developed to eliminate the time consuming step of cloning 5'RACE-PCR products prior to sequencing. Consequently, it allows for the rapid analysis of fusion transcript sequences from a large number of gene trap cell lines (Townley *et al.*, 1997). Moreover, in gene trap cell lines containing more than a single fusion transcript, direct sequencing produces multiple sequences superimposed on one another corresponding to the population of fusion transcripts present (Townley *et al.*,1997). The products of 5'RACE-PCR (Protocol 1) from 11.5 d.p.c. I114

homozygous liver RNA were analysed using the direct sequencing protocol. Figure 4.3 shows the autoradiograph of the direct sequencing products. The *en-2* exon sequence up to the splice acceptor site is easy to read as it represents a single sequence. Upstream of the splice acceptor site the sequence becomes more difficult to read because there is more than one endogenous sequence correctly spliced to *en-2*. However, a dominant sequence is identifiable and corresponds to the most abundant fusion transcript (Group I) identified from the cloning of RACE products. This pattern is characteristic of gene trap cell lines producing multiple fusion transcripts (Townley *et al.*, 1997) and substantiates what we had observed from the cloning of 5'RACE-PCR products from the I114 cell line.

The isolation of multiple fusion transcripts by RACE-PCR is inconsistent with the identification of only a single fusion transcript by Northern analysis. This presumably reflects the sensitivity of RACE-PCR at identifying transcripts expressed at levels undetectable by Northern analysis. If the proportion of the different fusion transcripts identifed after PCR amplification reflects their relative abundance in I114 tissues, then the Northern analysis is likely to be detecting expression of the abundant Group I fusion transcript.

4.2.2. Analysis of RACE Clone Sequence

The gene trap vector pT1-ATG allows for vector insertion events into the untranslated regions (UTRs) of transcripts to be recovered (Hill and Wurst, 1993). The presence of the *lacZ* start codon allows for protein translation and production of β-gal activity from UTRs. Translation of the four fusion transcript sequences in frame with the *lacZ* open reading frame (ORF) is shown in Figure 4.2. The Group I, III and IV fusion transcripts contain stop codons in frame with the ORF of the *lacZ* gene. Theoretically, if functional LacZ protein is produced from any of these fusion sequences, this would indicate that the vector is splicing to the 5'UTR of the endogenous gene and using the *lacZ* start codon. The endogenous sequence of the

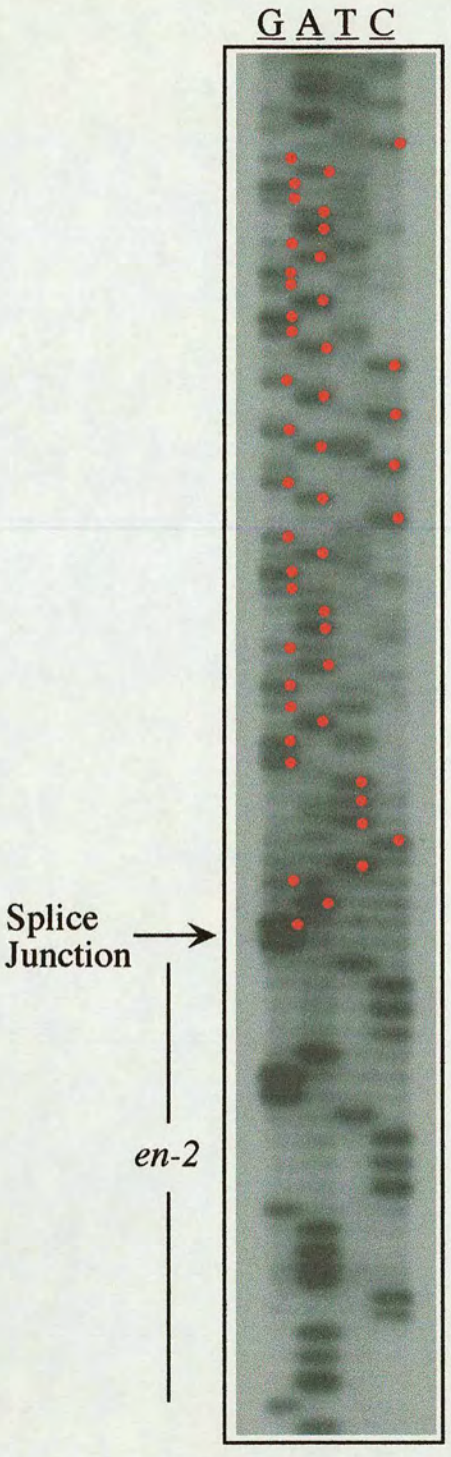


Figure 4.3: Direct Sequence of I114 fusion transcripts

Autoradiograph of direct sequencing products from I114 (-/-) 11.5dpc liver RNA separated on polyacrylamide gel. The en-2 exon sequence of PT1-ATG produces a clear, single sequence ladder. After the splice junction, the sequence is more ambiguous reflecting the multiple endogenous fusion sequences splice correctly to en-2 exon splice acceptor. The red dots highlight the strongest sequence which corresponds to the abundant Group I fusion transcript from RACE cloning.

Group II fusion transcript has no stop codons in-frame with the ORF of *lacZ*. Assuming the presence of a upstream start codon, translation of this sequence would produce active LacZ protein.

Endogenous sequence derived from the RACE clones of Groups 1-4 was compared to a non-redundant GenBank database using the BLASTN algorithm (Altschul *et al.*, 1990; National Center for Biotechnology Information (USA) website http://www.ncbi.nlm.nih.gov/index.html). The Group I sequence showed the highest level of homology with 86% identity over 46 nucleotides to the chick winged helix protein CWH-1 (accession number U37272; Freyaldenhoven *et al.*, 1997) with a probability of $1.4x10^{-5}$ of this match occuring by chance. Moreover, the same region of the Group I sequence showed comparable levels of homology to a number of genes from a range of species (data not shown). In all the genes showing homology to the Group I sequence, the region of homology translates to a negatively charged protein domain consisting of aspartate and glutamate residues (the exact order of the D and E residues is not maintained between genes). Translation of the Group I sequence in frame 2 produces a comparable stretch of aspartate and glutamate residues although, as with the other genes, the absolute order of these residues is not homologous between genes (Figure 4.2). The possible function of this negatively charged domain is discussed in Chapter 6. None of the other fusion sequences generated a significant level of homology to database sequences.

4.3. Expression Analysis of Group I and II Fusion Transcripts

From the RACE-PCR analysis, Group I and II fusion transcripts represent the only endogenous sequences isolated from <u>both</u> I114 ES cells and foetal liver. Groups III and IV were only isolated once each from ES cells and are therefore considered less significant. As the PCR technique is particularly sensitive to contamination and amplification artefacts (Kwok *et al.*, 1989; Porter-Jordan and Garret, 1990), RNase protection assays (RPA) were performed using Groups I and II sequences to confirm

the presence of the fusion transcripts and to assess their expression profile in I114 embryonic tissues.

(a) RNase Protection Assay with Group I Sequence

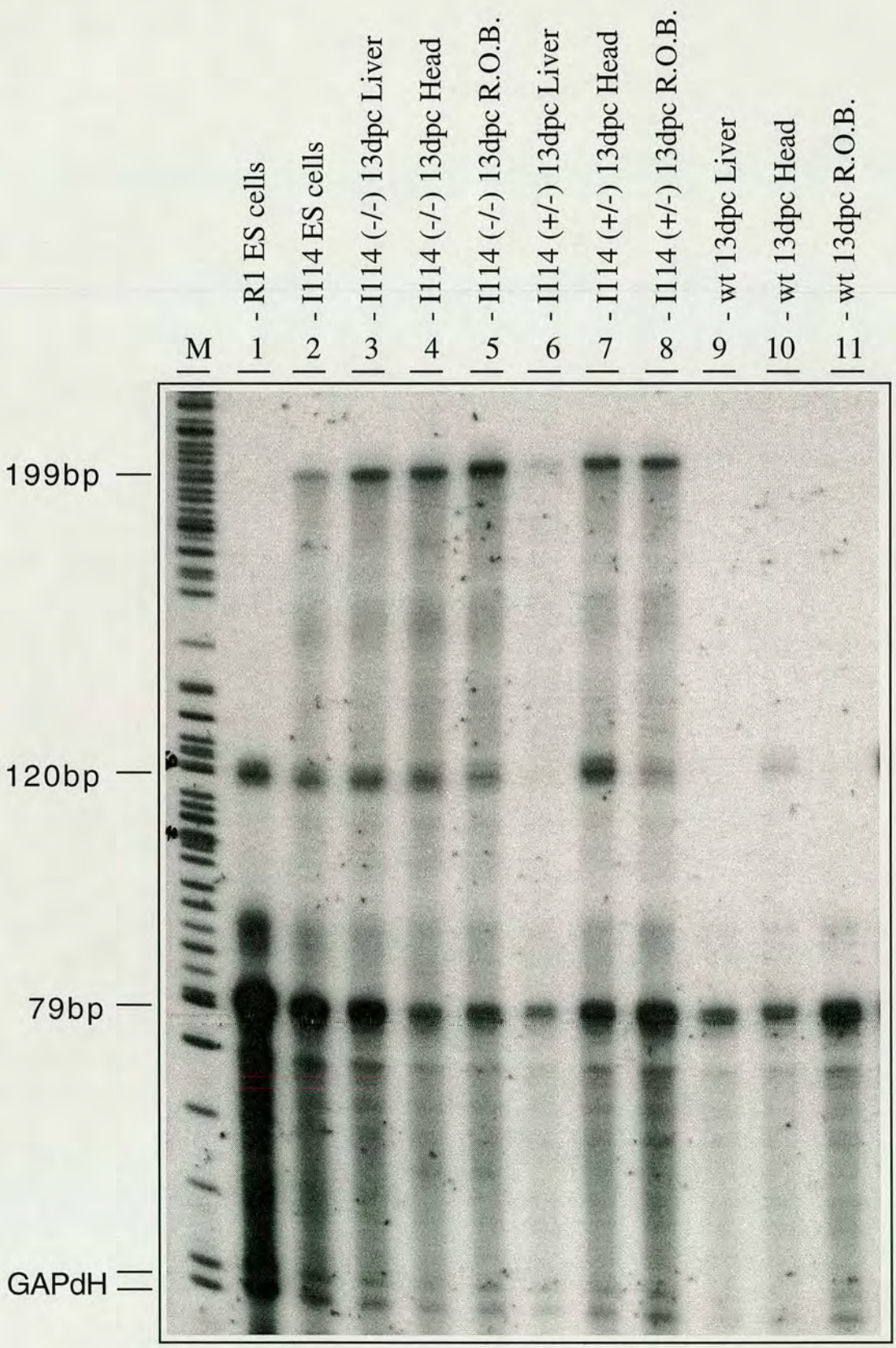
A ^{32}P-labelled riboprobe template was generated from clone L-A8 (Group I) in the antisense orientation to the fusion transcript. The L-A8 riboprobe was subsequently hybridised to RNA from the parental R1 ES cell line, I114 ES cells and liver, head and rest of body (R.O.B.) from wild type, I114 heterozygous and I114 homozygous 13d.p.c. embryos. GAPdH expression was used to control for RNA loading in each sample. Polyacrylamide gel electrophoresis of RNase digestion products resolves the fusion transcript as a 199bp protected fragment in all I114 tissues but not in wild type tissues (Figure 4.4, lanes 2 to 8). This ubiquitous expression pattern contradicts the I114 foetal liver specific embryonic β-gal activity. The protected fragment at 120bp corresponds to *en-2* transcript. In wild type tissues, *en-2* expression from the endogenous locus is observed solely in the embryonic head (Figure 4.4, lane 10) in agreement with the published expression pattern of *en-2* (Joyner and Martin, 1987). Expression of *en-2* is also observed in the parental R1 ES cell line (lane1). The same 120bp fragment is protected in all I114 tissues and corresponds to both endogenous *en-2* expression (in ES cells and head only) and expression of the *en-2* exon from the gene trap vector independent of this particular fusion transcript (Figure 4.4, lanes 2-8). The protected transcript of 79bp, observed at comparable levels in all tissues, corresponds to expression of the endogenous sequence independent of the fusion transcript (Figure 4.4). As with the full fusion transcript, its expression does not match I114 β-gal activity.

The expression of the Group I fusion transcript and endogenous sequence in all I114 embryonic tissues examined, namely the liver, head and rest of body, does not correspond to the liver specific expression of the reporter activity. There are two posssible explanations for this result. Firstly, the LacZ protein produced from the

Figure 4.4: Expression analysis of Group I fusion transcript LA-8 by RNase protection

A ^{32}P-labelled riboprobe was produced antisense to the LA-8 fusion transcript and hybridised to RNA from R1 and I114 ES cells, wild type, I114 homozygous and heterozygous liver, head and rest of body (R.O.B.) from 13d.p.c. embryos. After hybridisation, products were digested and run on a polyacrylamide gel.
The fusion transcript can be seen as a protected fragment of 199bp in all I114 tissues (lanes 2-8). *En-2* and endogenous transcripts, produced independently of the fusion transcript are protected as 120bp and 79bp fragments respectively. The GAPdH riboprobe is protected as 65 and 67bp fragments and controls for the amount of total RNA in the hybridisation reaction.
M; size marker, ddT terminated sequencing of -40 primed bacteriophage M13mp18 control DNA.

Figure 4.4:


M
1 - R1 ES cells
2 - I114 ES cells
3 - I114 (-/-) 13dpc Liver
4 - I114 (-/-) 13dpc Head
5 - I114 (-/-) 13dpc R.O.B.
6 - I114 (+/-) 13dpc Liver
7 - I114 (+/-) 13dpc Head
8 - I114 (+/-) 13dpc R.O.B.
9 - wt 13dpc Liver
10 - wt 13dpc Head
11 - wt 13dpc R.O.B.
199bp
120bp
79bp
GAPdH

ubiquitously expressed Group I fusion transcript is specifically translated and/or active in the embryonic liver. Alternatively, the Group I fusion transcript is not producing functional β-gal and an alternative fusion transcript, expressed exclusively in the foetal liver, is responsible for β-gal activity. This latter suggestion is supported by the isolation of three other fusion transcripts from the 5'RACE-PCR of I114 ES cell RNA. Moreover, the RPA shows expression of *en-2* independently of the fusion transcript in all I114 tissues (taking into account the low level endogenous expression in the head). This is indicative of the presence of other fusion transcripts containing the gene trap vector derived *en-2* exon.

The RPA using the Group I probe also shows that in I114 homozygous tissues, endogenous gene sequence is being produced independently of the Group I fusion transcript at levels approaching those seen in wild type tissues. This indicates that the endogenous gene may be splicing around the gene trap vector which could result in the production of wild type endogenous message, as has been observed for other gene trap lines (Table 1.1).

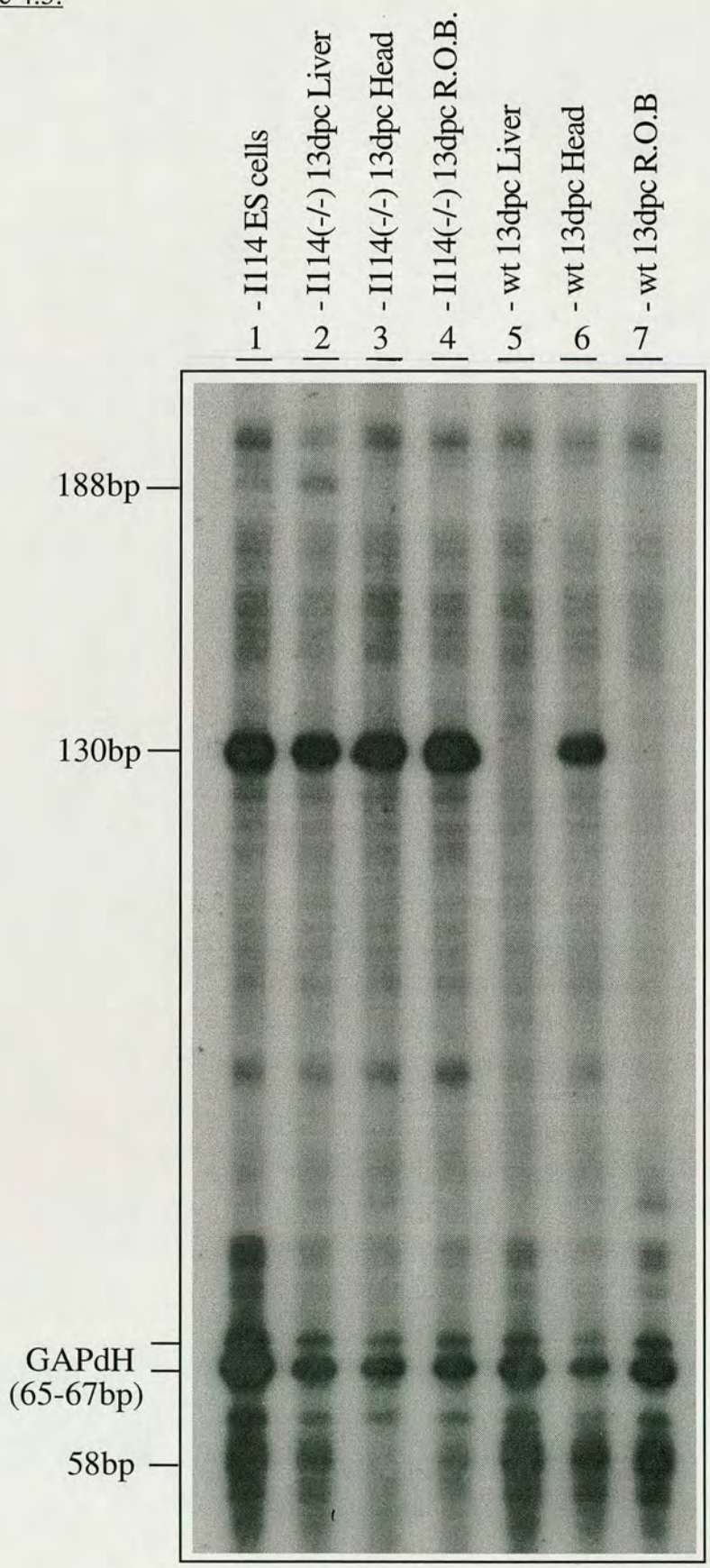
(b) RNase Protection Assay with Group II Sequence

An RPA was performed using an antisense riboprobe complementary to the L-69 (Group II) fusion transcript. This riboprobe was hybridised to RNA from I114 ES cells, I114 homozygous and wild type 13d.p.c. liver, head and R.O.B., digested and the products resolved on a polyacrylamide gel. The L-69 fusion transcript is protected as a 188bp transcript in I114 ES cells and 13d.p.c. I114 homozygous liver only and not in the head and R.O.B. (Figure 4.5). This expression pattern is comparable to the β-gal activity in the I114 line. As with the RPA shown in Figure 4.4, the *en-2* protected transcript (130bp) is present in all tissues derived from I114 embryos. This corresponds to *en-2* exon expression from the gene trap vector independent of this specific fusion transcript and will presumably include expression of the abundant

Figure 4.5: Expression analysis of Group II fusion transcript L-69 by RNase protection

A ^{32}P-labelled riboprobe antisense to L-69 was hybridised to RNA from I114 ES cells (lane 1) and liver, head and R.O.B. from I114 homozygous (lane 2, 3 and 4) and wild type 13d.p.c. embryos (lanes 5, 6 and 7). Products were digested, and resolved on a polyacrylamide gel. The fusion transcript is protected as a fragment of 188bp in I114 ES cells (lane 1) and I114 homozygous 13d.p.c. livers (lane 2). Expression of *en-2* independent of the fusion transcript is seen as a 130bp protected fragment in all I114 tissues (lanes 1 to 4) and in wild type head (lane 6). Expression of the endogenous sequence independent of the gene trap vector protects a fragment of 58bp in all samples except I114 homozygous 13d.p.c. head (lane 3).
GAPdH is protected as a 65bp and 67bp fragment and controls for the amount of total RNA used in the hybridisation reaction.

Figure 4.5:


1 - I114 ES cells
2 - I114(-/-) 13dpc Liver
3 - I114(-/-) 13dpc Head
4 - I114(-/-) 13dpc R.O.B.
5 - wt 13dpc Liver
6 - wt 13dpc Head
7 - wt 13dpc R.O.B
188bp
130bp
GAPdH
(65-67bp)
58bp

(GroupI) fusion transcript. In wild type tissues, *en-2* expression is seen exclusively in the foetal head as predicted from the expression of endogenous *en-2*.

Expression of the endogenous L-69 sequence independent of the fusion transcript is protected as a 58bp fragment (Figure 4.5). Expression can be seen in I114 ES cells, I114 homozygous foetal liver and R.O.B.and all tissues of wild type embryos (lane 1, 2 and 4-7, Figure 4.5). This result is difficult to interpret. Expression of Group II endogenous sequence has since been shown to be liver specific in tissues from wild type embryos (Chapter 6). Although there is a protected fragment at the expected size for the endogenous sequence, the resolution around this region is poor with several bands present. These bands may represent digestion artefacts from the larger protected fragments, for example GAPdH, and not specifically expressed products.

(c) RT-PCR Analysis of Group II Sequence

To confirm the results of the Group II RPA in embryonic tissues and to extend the expression analysis to adult I114 tissues, RT-PCR was performed using primers complementary to the L-69 (Group II) fusion transcript. Primer LST-1, derived from clone L-69 endogenous sequence and primer 78, complementary to the 5' end of the *lacZ* gene were designed and used in the PCR reaction (Figure 4.6A+B). This is predicted to amplify a PCR product of 377bp in tissues expressing this transcript. HPRT primers were used to control for the amount of total RNA in each reaction (Johansson and Wiles, 1995). The PCR products were separated on an agarose gel, Southern blotted and probed with the *en-2* exon probe (Figure 4.6B). The L-69 fusion transcript is amplified from I114 ES cell (lane 1) and 13d.p.c. I114 homozyous liver RNA (lane 2) with no significant amplification product observed from I114 homozygous head (lane 3). A faint band is observed in the R.O.B. RNA sample (lane 4) which is likely to be a result of contamination from the incomplete removal of liver tissue from the rest of body prior to RNA preparation. From the adult RNA samples, expression of the L-69 fusion transcript is seen at highest levels in the ovaries (lane 6)

Figure 4.6: Expression Analysis of Group II Fusion Transcript L-69 by RT-PCR

A. Sequence of L-69 fusion transcript showing the primer LST-1 used in the RT-PCR experiment. Lower case letters show *en-2* exon sequence.
T(blue); poly T tail added artificially during RACE-PCR (protocol 2).

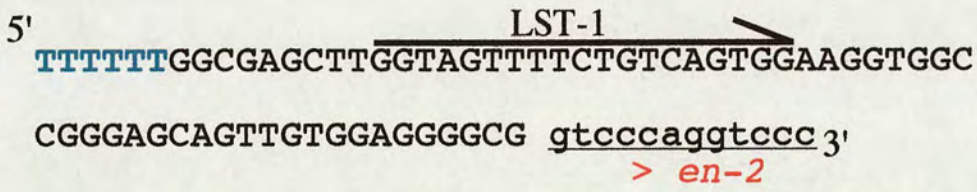
B. Predicted fusion transcript showing primers and probe used in RT-PCR assay.

C. Results of RT-PCR using the LST-1 primer from L-69 and primer 78 from *lacZ*.

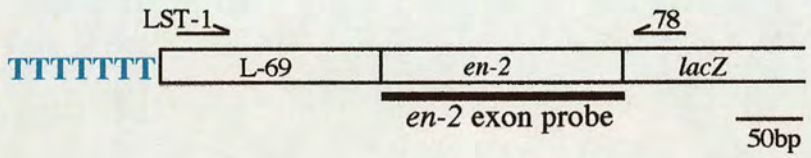
RT-PCR amplification products using primers LST-1 and 78 were separated on an agarose gel, Southern blotted and probed with ^{32}P-radiolabelled *en-2* exon. The expected 377bp product is observed at its highest level in I114 ES cells, I114(-/-) 13d.p.c. liver and I114 (-/-) adult ovaries and kidney (lanes 1, 2, 6, and 11 respectively). Lower levels are seen in I114(-/-) 13d.p.c. R.O.B. and I114(-/-) adult lung, liver and heart (lanes 4, 8, 9 and 10 respectively). Lane 13 shows I114(-/-) liver RNA subjected to the RT-PCR protocol in the absence of reverse transcriptase. Amplification of HPRT is used to control for the amount of total RNA used in the RT reaction. The presence (+) or absence (-) of β-gal enzymatic activity is indicated for comparison.

Figure 4.6:

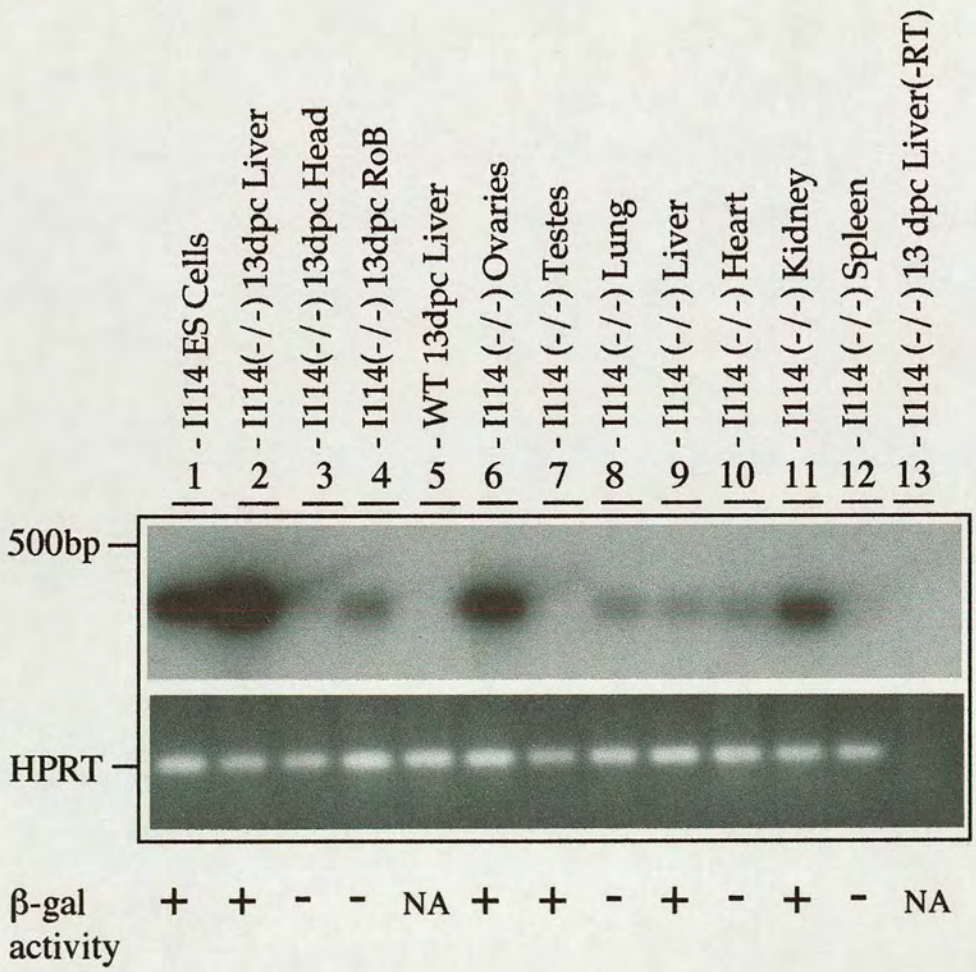
A: L-69 - 58bp



B: RT-PCR strategy



C: RT-PCR



and kidney (lane 11), with low levels seen in the lung (lane 8), liver (lane9) and heart (lane 10). No product was amplified from wild type 13d.p.c. foetal liver (lane 5) eliminating the possibility that the product amplified from I114 homozygous foetal liver is specific to foetal liver RNA rather than the I114 line . As the genomic structure of the vector integration is unknown it is important to control for potential genomic DNA contamination of the I114 homozygous foetal liver RNA sample. When reverse transcriptase was omitted from the RT-PCR reaction, no product was amplified (lane 13) indicating no genomic DNA contamination.

In summary, the expression of L-69 fusion transcript (Group II) in 13d.p.c. I114 homozygous embryos correlates with the embryonic profile of I114 β-gal activity. In the adult, the highest level expression of the L-69 transcript is seen in the kidney and ovary correlating with β-gal activity in these tissues (Figure 3.3). Lower level expression is seen in the liver, lung and heart. Although reporter activity is seen in the adult liver, none has been identified in the lung or heart. This may reflect L-69 expression in the lung and heart being sensitive to detection by RT-PCR but not of a sufficient level to confer detectable β-gal activity. The L-69 fusion transcript was not detected in the adult testes where reporter expression is observed. The RT-PCR protocol was repeated using different RNA preparations of I114 homozygous testes and kidney. Using the kidney sample as a positive control, a low level of L-69 transcript was amplified in the testes (data not shown).

Both the RT-PCR and RPA studies of embryonic tissues show that expression of the L-69 fusion transcript is highly restricted to the liver at mid gestation, correlating well with I114 reporter activity. In addition, RT-PCR analysis of L-69 expression in adult tissues correlates well with adult I114 reporter activity. This data suggests that the L-69 fusion transcript is directly responsible for the β–gal activity observed in I114 embryos and adult.

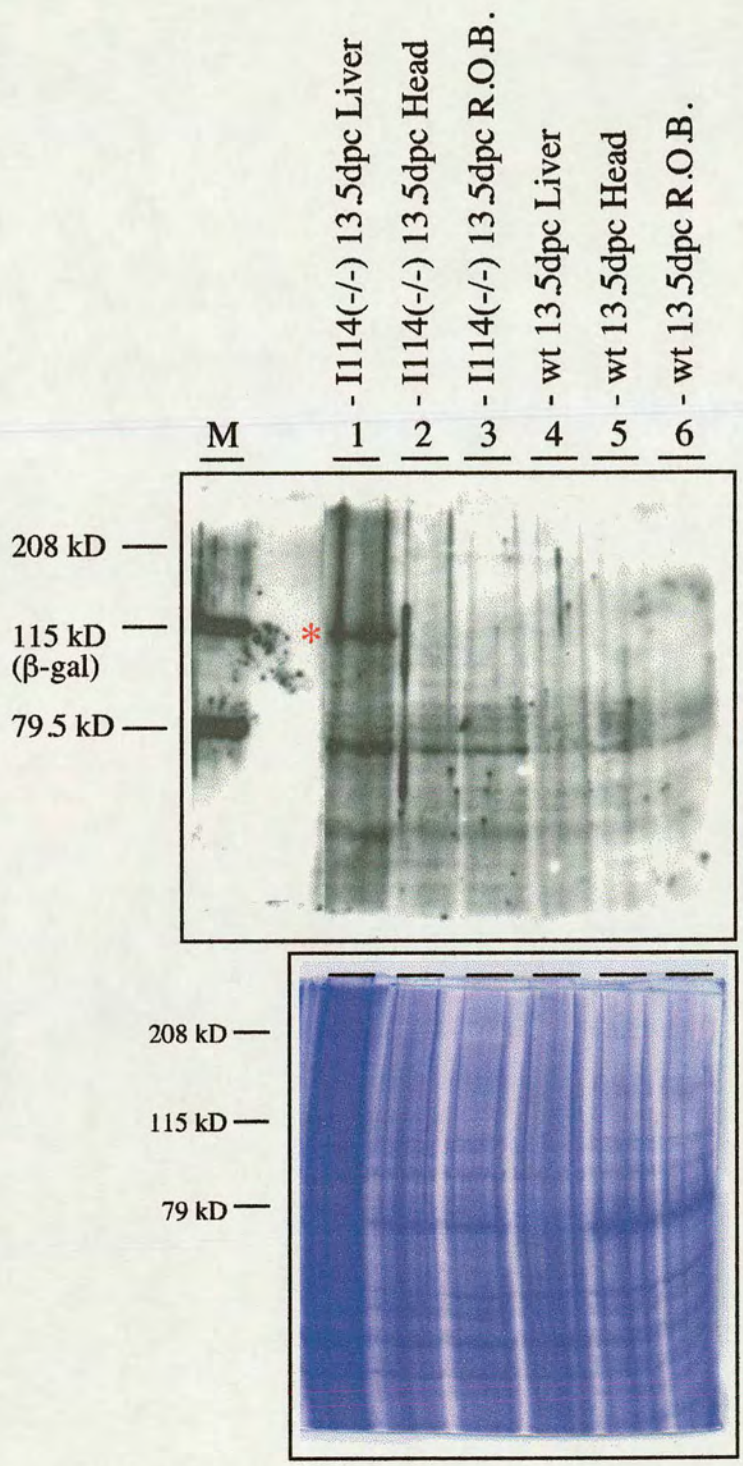
4.4. LacZ Protein Expression Analysis

It is important to determine whether expression of the β-gal protein is restricted to the foetal liver or whether its enzymatic activity is restricted to the liver of I114 embryos with the I114 gene trap integration expressing a predominant fusion transcript in all embryonic tissues. Crude protein lysates (20μg) from 13.5d.p.c. wild type and I114 homozygous liver, head and R.O.B. were separated by SDS-polyacrylamide gel electrophoresis (PAGE), Western blotted and probed with a primary polyclonal (rabbit) anti-β-gal rabbit antibody. The primary antibody was probed with Horseradish peroxidase-labelled anti-rabbit Ig. antibody conjugate followed by chemiluminescent detection. The same amount of protein used in the Western blot was separated by SDS-PAGE and stained with Coomassie Blue to control for protein loading. Despite the uneven loading of the protein samples, it is apparent that only the I114 homozygous liver (lane 1) expresses the 115kD β-gal protein (Figure 4.4). The wild type samples serve as controls for the possibility of non-specific antibody binding. This result shows that the restricted β-gal activity in the foetal liver is a result of restricted protein expression excluding the possibility that β-gal protein is expressed throughout the embryo but only enzymatically active in the foetal liver. Therefore, the I114 liver specific β-gal activity is a result of either liver specific translation or transcription. Within the sensitivity and resolution afforded by the Western blot analysis, the size of the β-gal protein in the I114 foetal liver suggests that it does not exist as a large fusion protein. This is expected from the size of the I114 fusion transcript identified by Northern analysis in Figure 4.1.

Figure 4.7: β-galactosidase protein expression in I114 embryos

Western blotting and Coomassie staining of protein samples from I114 homozygous and wild type 13d.p.c. liver, head and R.O.B after separation by SDS-PAGE.
The upper panel shows a Western blotted gel which has been probed with mouse polyclonal anti-β-gal antibody. Anti-rabbit Ig antibody conjugated to horseradish peroxidase was then used to probe for primary antibody which was in turn visualised by chemiluminescence.
Only the I114(-/-) 13d.p.c. liver contains the β-gal protein at the expected size of 115kD (lane 1-asterisk).
M; Molecular Weight Standards (Biorad, Cat No. 161 0318). 115kD standard corresponds to β-galactosidase. The positive size marker at 79.5 kD suggests that bovine serum albumin is cross reacting with either the primary anti-β-gal antibody or the secondary anti-rabbit antibody.
The lower panel shows a second gel, with the same amount of protein as the Western blotted gel, after Coomassie Blue staining which serves as a protein loading control.

Figure 4.7:


M
1 - I114(-/-) 13.5dpc Liver
2 - I114(-/-) 13.5dpc Head
3 - I114(-/-) 13.5dpc R.O.B.
4 - wt 13.5dpc Liver
5 - wt 13.5dpc Head
6 - wt 13.5dpc R.O.B.
208 kD
115 kD
(β-gal)
79.5 kD
208 kD
115 kD
79 kD

4.5. Discussion

The molecular characterisation of the I114 gene trap integration has identified a complex integration event with multiple fusion transcripts isolated from both ES cells and foetal liver. The most abundant Group I fusion transcript is expressed throughout the embryo while expression of the rarer Group II fusion transcript correlates well with the highly restricted β-gal activity in the embryo and adult.

The production of multiple fusion transcripts in individual gene trap cell lines has been reported previously. From the RACE-PCR cloning of 55 gene trap cell lines, 3 lines produced multiple fusion transcripts (Chowdhury *et al.*, 1997). The direct sequencing of 153 secretory trap cell lines identified 14 lines (9%) with multiple sequences correctly spliced to the gene trap vector (Townley *et al.*, 1997). In both the studies, these integrations were not characterised further. Certain gene trap vector integration events can be predicted to result in the production of multiple fusion transcripts. These include the integration of multiple vectors into several sites in the genome, *trans*-splicing and alternative splicing from a single vector integration site to more than one exon (Figure 4.8).

The I114 gene trap integration contains a minimum of four gene trap vector copies. It is therefore possible that individual vectors have inserted into transcription units at separate chromosomal sites where splicing to different upstream exons occurs (Figure 4.8A). Five generations of backcrossing the I114 gene trap integration onto the C57BL/6 genetic background has failed to segregate any of the vector copies, indicating some degree of linkage between the vector copies. Interestingly, in the cell lines I23 and I163, identified in the same RA pre-screen as I114 and also containing multiple vector copies, a single vector copy was segregated after only two generations of backcrossing (Forrester *et al.*,1996). Chapter 5 describes experiments which define more accurately the linkage between the vector copies in the I114 gene trap integration.

The study of two separate gene trap integrations into RNA polymerase I transcribed rRNA genes identified that such insertions induced the *trans*-splicing of the

Figure 4.8: Mechanisms of multiple fusion transcript production

A: Vector insertion into separate chromosomal sites.
Insertion of the gene trap vector into different transcription units which could be located relatively closely to one another or on different chromosomes.

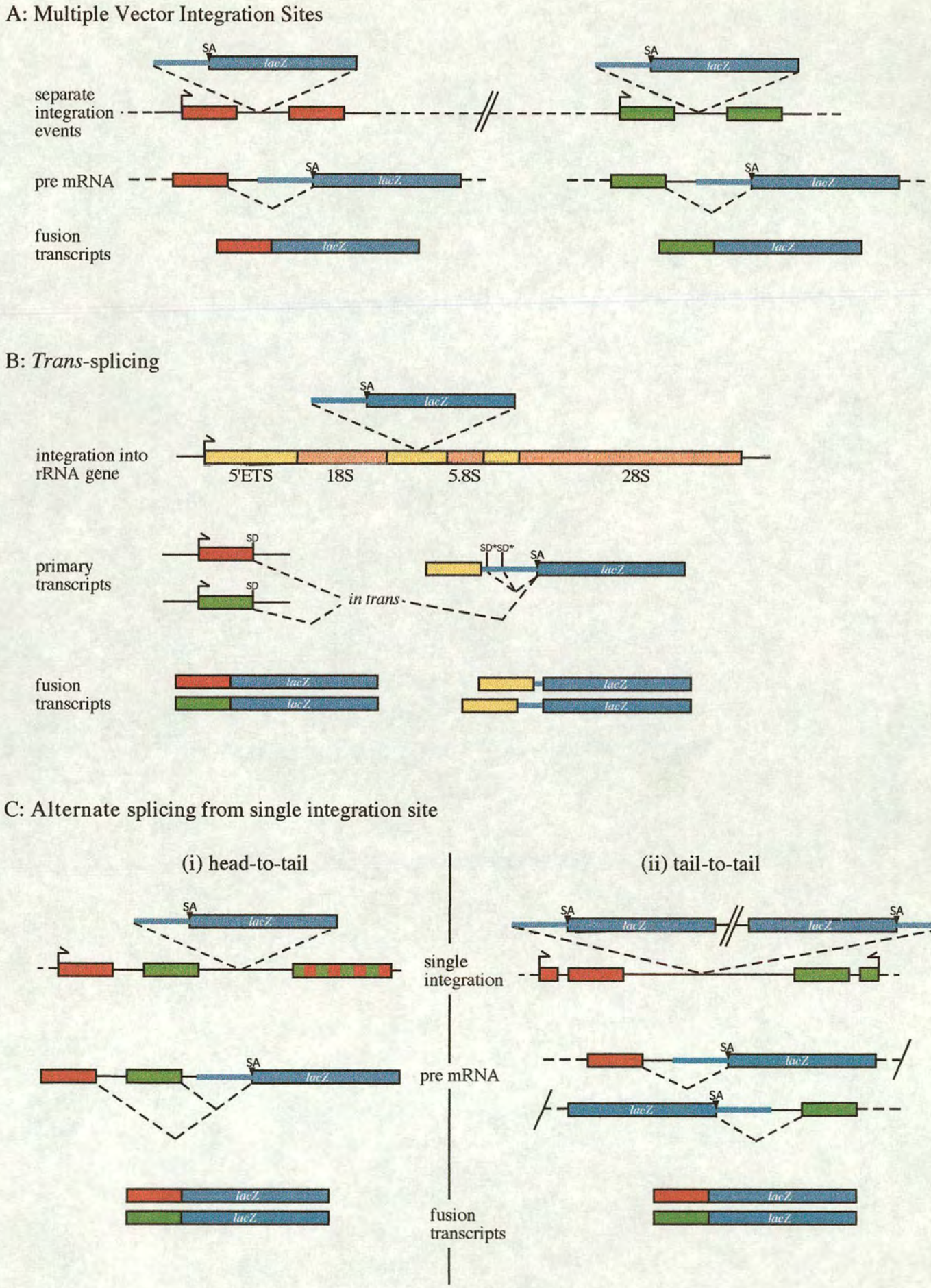
B: *Trans*-splicing after insertion into RNA polymerase I and III transcription units.
Insertion of a gene trap vector into the extra transcribed spacer (ETS-yellow boxes) of the rRNA gene will produce a pre-rRNA with the gene trap vector downstream of the ETS. This results in the splicing of the vector splice acceptor to *in cis* to two cryptic splice sites within the vector intron or *in trans* to endogenous splice donor (SD) sequences. This will produce fusion transcripts containing endogenous sequence and rRNA/*en-2* intron sequence. As the rRNA transcripts are uncapped, they will not be substrates for translation.

C: Alternate splicing of gene trap vector after insertion into a single chromosomal site.

(i) Tandem insertion of the gene trap vectors in a head-to-tail orientation. Vector splicing to alternate upstream endogenous exons would produce multiple fusion trascripts.
(ii) Tandem insertion of the gene trap vector in a tail-to-tail orientation. Vector splicing to alternate exons from two different converging transcripts

SA, vector splice acceptor; SD, endogenous splice donors; SD*, cryptic splice donors in the *en-2* intron sequence at positions -32 and -435; arrows=direction of transcription.

Figure 4.8:


A: Multiple Vector Integration Sites
separate integration events
pre mRNA
fusion transcripts
SA
lacZ
B: Trans-splicing
integration into rRNA gene
5'ETS
18S
5.8S
28S
primary transcripts
SD
SD*SD*
in trans
fusion transcripts
C: Alternate splicing from single integration site
(i) head-to-tail
(ii) tail-to-tail
single integration
pre mRNA
fusion transcripts

gene trap vector to several different endogenous gene sequences (J. E. Sleeman & W. C. Skarnes, unpublished results). The exact mechanism of *trans*-splicing is unclear. One hypothesis is that the low efficiency of *cis*-splicing by RNA polymerase I transcripts allows the interaction of splicing machinery associated with the vector splice acceptor to interact with other splice donors *in trans* (W. C. Skarnes, personal communication; Figure 4.8B). In both these *trans*-spliced lines, fusion transcripts containing *en-2* intron sequence were also isolated. However, these transcripts did not represent unspliced *en-2* intron sequence. Rather, the production of the entire *en-2* intron-exon sequence as a part of the RNA polymerase I transcript was predicted to have induced the splicing of the *en-2* splice acceptor to two upstream cryptic splice donor sequences in the *en-2* intron sequence (Figure 4.8B). Although the *en-2* intron sequences isolated from the I114 gene trap cell line represent unspliced intron, *trans*-splicing could still be a feature of the I114 gene trap integration.

Another possibility is that the I114 gene trap vectors have integrated into a single chromosomal site where they interact with multiple splice donor sites (Figure 4.8C). Integration of the vectors, such that the two external vector copies are in a head-to-tail tandem array within a single transcription unit, could result in vector splicing to alternative upstream exons (Figure 4.8C(i)). Vector insertion downstream of exons, whose splicing is normally regulated in a tissue specific manner (for examples see Adams *et al.*, 1996), could produce alternate splicing. Alternatively, the presence of the splice acceptor of the gene trap vector could disrupt the normal splicing pattern of the endogenous gene and induce alternate splicing. Another possibilty is that the integration of the external vector copies in a tail-to-tail orientation between two adjacent genes transcribed in opposite directions could present two separate vector splice acceptor sequences to different upstream exons (Figure 4.8C(ii)). A similar situation could potentially arise from the insertion of a head-to-head vector array in the same site.

As well as identifying different fusion transcripts, *en-2* intronic sequence has repeatedly been isolated by RACE-PCR from different I114 tissues. Moreover, *en-2* intron has also been isolated from the screening of an I114 homozygous 13d.p.c. liver

cDNA library (constructed by Pierre Drèze, Brussels) with an *en-2* exon probe (data not shown). It is perhaps more likely that the isolation of this sequence reflects inefficient vector splicing rather than an artefact arising from genomic DNA contamination of the RNA samples. In the study of Townley *et al.*, (1997), 21 (14%) of the 153 directly sequenced cell lines produced *en-2* intronic sequence with endogenous fusion sequence. Moreover, all these lines were subsequently shown to express *en-2* intron sequence by Northern blot analysis. The failure to detect *en-2* expression by Northern blot in I114 ES cells probably reflects the low level of *en-2* intron expression these cells.

Inefficient vector splicing has been hypothesised to be a feature of several different gene trap vector integration events. The insertion of a gene trap vector into the exon of an endogenous gene is predicted to result in *en-2* intron expression. Such an event would result in competition between the vector splice acceptor and the endogenous splice acceptor of the exon into which insertion has occurred (which will be immediately upstream). If splicing to the endogenous splice acceptor occurred, a transcript containing the vector intronic sequence as part of the endogenous exon would be produced. From several cell lines characterised in our own and Bill Skarnes laboratory, it appears that insertion into an endogenous exon induces splicing between the *en-2* splice acceptor and cryptic splice donor sequences within the *en-2* intron as seen for *trans*-spliced lines (McClive *et al.*,1998; W. C. Skarnes, unpublished results). Moreover, this cryptic splicing was relatively efficient as no unspliced vector sequences were isolated from the RACE cloning of two of these lines (McClive *et al.*, 1997).

Another possible cause of inefficient splicing is that it is a intrinsic property of the site into which the vector has integrated. For example, if vector insertion has occurred into a large intron, the distance between the vector and endogenous exon may affect splicing efficiency.

How do Multiple Fusion Transcripts Produce Specific Reporter Activity?

The presence of the *lacZ* start codon in pT1-ATG means that all of the fusion transcripts isolated so far could potentially produce active β-gal. How does the I114 gene trap cell line display such a localised reporter expression pattern in the embryo when multiple fusion transcripts are produced, the most abundant of which is ubiquitously expressed? The Western blot analysis shows that the β-gal protein expression is restricted to the foetal liver. It is therefore likely that only the liver specific fusion transcript is translated and produces β-gal activity. From the translation of the different fusion transcripts, only the liver specific fusion transcript (Group II, Figure 4.2) has no stop codons in-frame with the *lacZ* ORF. Moreover, the prevalent fusion transcript (Group I) and the Group IV fusion transcript contain at least one frame with no stop codons when translated in all three frames. Translation in this frame may correspond to the open reading frame of the endogenous sequence, which, in the context of splicing to the gene trap vector, would place *lacZ* out of frame. The Group III fusion transcript contains stop codons in all reading frames and should therefore translate *lacZ* from its own start codon. However, both Group III and Group IV transcripts have never been isolated from foetal liver RNA and only a single clone of each was isolated from ES cells making it unlikely that they are expressed at sufficiently high levels to produce detectable β-gal activity. One way of resolving this question would be to isolate the corresponding endogenous gene sequence of each fusion transcript, conceptually translate these sequences and analyse if the fusion transcripts are within the hypothesised ORF.

An interesting aspect of this situation is that, although potentially not responsible for the β-gal activity, the wild type message of the Group I fusion transcript should be disrupted by the gene trap vector. Consequently, any phenotype(s) arising from the I114 gene trap integration could be attributed to the disruption of the prevalent fusion transcript and/or the disruption of the other fusion transcripts. However, from the RNase protection assay using the Group I fusion transcript L-A8 as a probe, it is

apparent that animals homozygous for the I114 gene trap integration produce significant levels of endogenous message. This suggests that the endogenous gene is splicing around the gene trap vector and therefore is not a null allele. The same situation may also be occuring with the Group II liver specific fusion sequence although expression of the endogenous transcript in homozygous tissue is unclear from the L-69 RPA data.

In the I114 gene trap cell line, the Group II fusion transcript is likely to be responsible for the β-gal activity observed in the embryo and adult. The repeated screening of a 12-13d.p.c. wild type foetal liver cDNA library (constructed by Pierre Drèze, Brussels) with the fusion transcript L-69 (Group II) has failed to isolate the endogenous cDNA corresponding to this sequence. This prevents the analysis of this sequence independent of the gene trap vector to confirm the liver specific expression pattern. Even with such striking reporter activity, the possibility remains that this activity may be an artefact of the gene trap vector integration and not represent an endogenous transcript expressed exclusively in the foetal liver. The failure to isolate Group II sequence from this foetal liver library may reflect the low level of expression at this stage in gestation. The screening of genomic DNA for Group II sequence would provide a means of isolating this sequence independently of its expression state.

In Chapter 5, the liver specific fusion sequence is isolated from genomic DNA independently of the gene trap vector. The relationship between this liver specific fusion sequence (Group II) and the Group I transcript is subsequently determined at the genomic level helping to define an alternative splicing event resulting in the production of the Group I and II fusion transcripts.

Chapter 5

RESULTS

Genomic Characterisation of the I114 Gene Trap Integration

5.1 Introduction

As stated in the previous chapter, the I114 gene trap integration produces multiple fusion transcripts which we predicted to be a consequence of either individual vector insertion at separate chromosomal sites, *trans*-splicing or alternate splicing from the vectors integrated at a single chromosomal site (Figure 4.8). The work described in this chapter sets out to define the relationship between the two major fusion transcripts, the ubiquitous Group I and the liver specific Group II.

5.2. Chromosomal Location of the I114 Gene Trap Integration

To examine if the gene trap vectors have integrated into more than one chromosomal site and to identify the chromosomal location of the vector copies, pT1-ATG was used as a probe for fluorescent *in situ* hybridisation (FISH) analysis of chromosomal spreads from I114 ES cells (Muriel Lee, MRC HGU). Within the resolution afforded by FISH analysis, the vector copies resolved to a single chromosomal site which the G-banding pattern of the chromosome suggested was chromosome 5, band E1 (data not shown). The chromosomal location was confirmed from the hybridisation of texas red labelled pT1-ATG in parallel with a chromosome 5 specific fluorescein isothiocyanate (FITC) paint (Figure 5.1A; Rabbits *et al.*, 1995).

A total of 8 cDNAs were isolated using the Group I fusion transcript L-A8 (see Chapter 6, Appendix III). We used the largest of these clones GC7 (5.1Kb) for the

Figure 5.1: Chromosomal mapping of the I114 gene trap integration

A: PT1-ATG

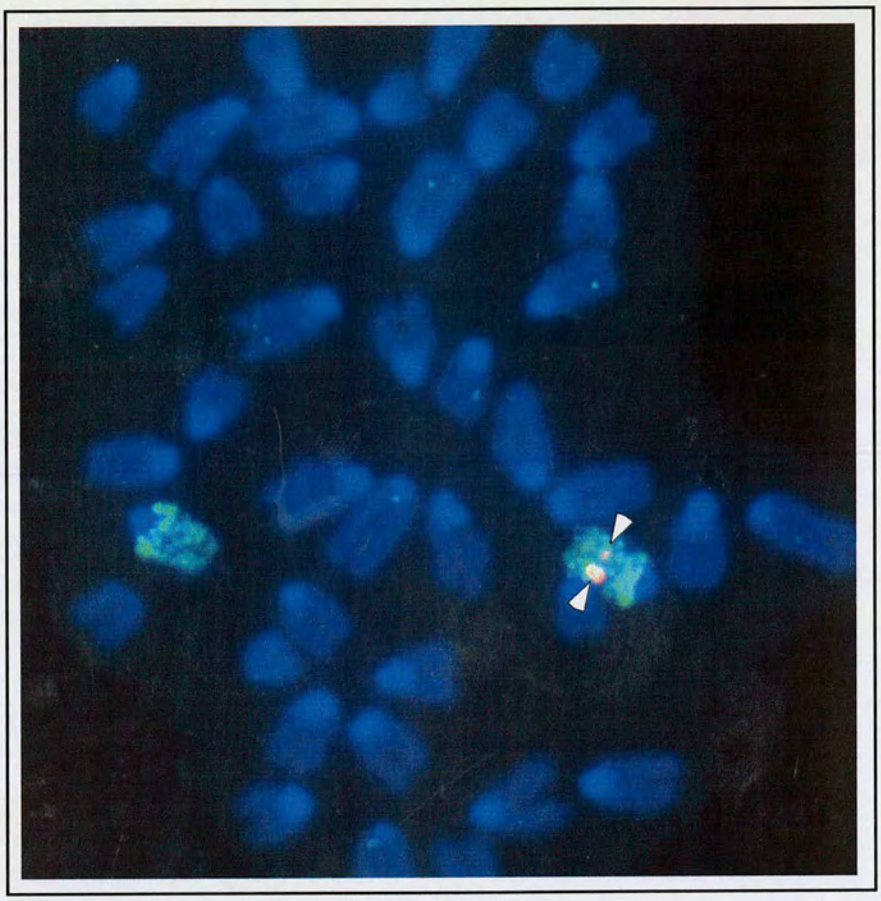
FISH analysis of I114 ES cell metaphase chromosomes with flourescently labelled pT1-ATG. Green flourescence identifies chromosome 5 (FITC Ch5 specific paint). pT1-ATG (texas red labelled) hybridises to a single chromosomal site on chromosome 5 (arrowed).

B: PT1-ATG and GC7

FISH analysis of I114 ES cell metaphase chromosomes with FITC (green) labelled PT1-ATG and texas red labelled GC7. Co-localisation of hybridisation of PT1-ATG and GC7 resolves as yellow flourescence (arrowed).

Figure 5.1:

A.



B.

FISH analysis of wild type cells and showed that it mapped to the same chromosomal location as the gene trap vector (data not shown)

Two colour FISH analysis was performed on I114 ES cell chromosome spreads using FITC labelled PT1-ATG and texas red labelled GC7 as probes. Figure 5.1B shows the colocalisation of PT1-ATG (green) and GC7 (red) on I114 ES cell metaphase choromosome spreads. This data shows that the endogenous gene associated with the Group I fusion transcript and the gene trap vector map to the same chromosomal site. The Group II fusion transcript has not been mapped by FISH due to the failure to isolate an associated cDNA clone.

5.3. Genomic Structure Analysis of the I114 Vector Copies

To define more accurately the proximity of the gene trap vector copies to each other, I114 ES cell genomic DNA was digested with rare cutting restriction enzymes in an attempt to resolve the multiple copies of the gene trap vectors to a single genomic DNA fragment. Digestion products were separated by pulse field gel electrophoresis (PFGE), Southern blotted and hybridised to a ^{32}P-radiolabelled *lacZ* probe. Digestion by both SalI and XhoI, which do not cut within pT1-ATG, resolves all the gene trap vector copies to a single genomic restriction fragment of approximately 145kb (Figure 5.2). This data strongly suggests that the gene trap vector copies have integrated into a single genomic site as a concatemeric array.

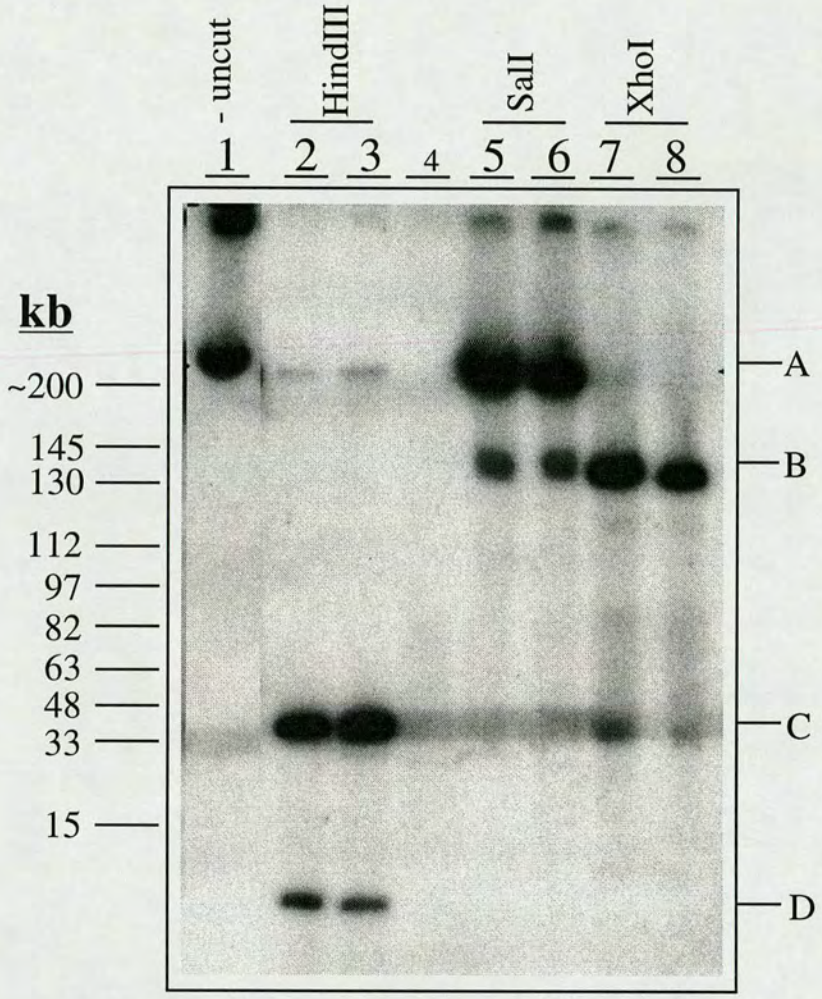
HindIII digestion of I114 genomic DNA resolves two positively hybridising fragments at approximately 33Kb and 10Kb (Figure 5.2). The fragment intensities suggest that the 33Kb fragment represents 3 copies of the gene trap vector and the 10Kb fragment a single vector copy. The production of the I114 gene trap cell line involved linearisation of the pT1-ATG vector prior to electroporation with HindIII. The separation of gene trap vector copies by a HindII probably reflects the production of a HindIII site after extrachromosomal ligation of the gene trap vectors and their

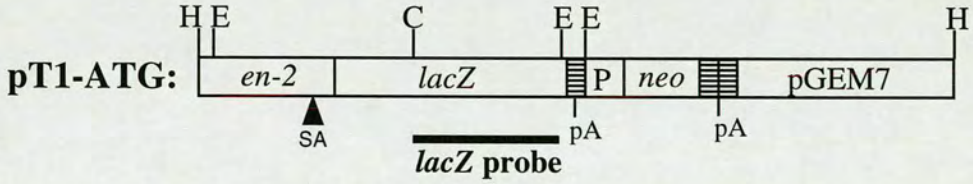
Figure 5.2: PFGE analysis of I114 genomic DNA

- The positively hybridising fragments at 200kb represent undigested genomic DNA (A). The uncut sample shows very little non-specific degradation with most of the *lacZ* positive genomic DNA greater than 200Kb in size. A large proportion of the *lacZ* positive genomic DNA in the SalI digest (lanes 5 & 6) remains undigested (>200Kb) suggesting that this has been an inefficient reaction.
- SalI and XhoI digested I114 genomic DNA (lanes 5-8) resolves all vector copies to a fragment of approximately 140kb (B).
- HindIII digested I114 genomic DNA (lanes 2 &3) identifies two positively hybridising fragments of approximately 33kb and 10kb (C+D).

- pT1-ATG: Basic restriction map of the gene trap vector showing the rare cutting restriction enzyme sites used in the study. The black bar indicates the *lacZ* fragment used to probe the Southern blot.
C, ClaI; E, EcoRI; H, HindIII; SA, splice acceptor; pA, polyadenylation signal.

Figure 5.2:


- uncut
HindIII
SalI
XhoI
1 2 3 4 5 6 7 8
kb
~200
145
130
112
97
82
63
48
33
15
A
B
C
D


H E
C
E E
H
pT1-ATG:
en-2
lacZ
P
neo
pGEM7
SA
pA
pA
lacZ probe

subsequent genomic integration rather than the separation of vector copies by genomic DNA.

5.4. Cloning and Analysis of Group I and II Genomic DNA

The mouse genomic P1 derived artificial chromosome (PAC) library RPC121 was constructed from female (129/SvEvTACfBr) spleen genomic DNA (K. Osoegawa & P. de Jongs, unpublished results; http://bacpac.med.buffalo.edu). This library is available from the UK HGMP Resource Centre on 7 on gridded filters (http://www.hgmp.mrc.ac.uk/Biology/descriptions/mouse_pac.html). The library contains 128,889 recombinant clones with an average insert size of 147Kbp which represents 6.3 times genomic coverage. Each clone has been spotted in duplicate onto the membranes, providing each with a unique co-ordinate which can subsequently be used to order these clones from the UK HGMP Resource Centre.

The Mouse PAC1-7 filters were initially probed with *en-2* exon sequence to exclude these clones from future screenings with *en-2* containing fusion transcript sequences. Six positively hybridising clones were identified, correlating well with the predicted 6.3 times genomic coverage of this library. The filters were stripped, probed with the Group II fusion transcript (L-69) and 3 positive clones were isolated (352-j9; 390-m1; 439-a23).

An oligonuleotide (UBT-1) complementary to the Group I fusion transcript (Figure 4.2) was used to probe Southern blotted DNA from the three Group II positive PAC clones. All three clones hybridised to UBT-1 (data not shown) indicating that the Group I and Group II sequences were directly linked at the genomic level.

Restriction digests of 439-a23 were carried out with individual or combinations of the rare cutting restriction enzymes ClaI, NotI, SalI, SfiI, XhoI. Digestion products were initially resolved by PFGE and subsequently by normal gel electrophoresis when restriction fragments were less than 15kb in size. The gels were Southern blotted and hybridised to a number of different probes. The L-69 fusion transcript and the UBT-1

oligonucleotide were used as Group I and II fusion sequence probes respectively. End labelled SP6 and T7 primers defined restriction fragments containing vector sequence and served to orientate the genomic DNA insert (Figure 5.3A). Both the liver specific Group II fusion sequence (L-69) and the Group I fusion sequence (UBT-1) mapped at the NotI/SfiI/XhoI sites (Figure 5.3A). A 1.5Kb PstI restriction fragment that hybridised to both Group I and Group II probes (Figure 5.3B) was subcloned into pZERO™-2 (Invitrogen) and sequenced (Appendix I).

Analysis of the sequence reveals that within this genomic DNA fragment both the Group I and Group II endogenous fusion sequences are in the same orientation relative to the splice donor sites used by the gene trap vector (Figure 5.3C; Appendix I). Consequently, a single gene trap vector would be capable of alternate splicing to the Group I and II fusion transcripts from a single chromosomal integration site.

The genomic sequence was entered into the NIX suite of DNA analysis programs available at the UK HGMP Resource Centre (http://www.hgmp.mrc.ac.uk). NIX provides a single interface with which to run multiple DNA analysis programs against a specified DNA sequence (Appendix II).

(a) CpG island

CpG islands have a high overall guanine and cytosine nucleotide content as well as a high frequency of CpG dinucleotides relative to bulk genomic DNA. CpG islands are associated with the 5' ends of housekeeping genes as well as the 5' and 3' of many tissue specific genes (Gardiner-Garden and Frommer, 1987). The GRAIL CpG island algorithm (using the definition of a CpG island as reported by Gardiner-Garden and Frommer, 1987) defines an excellent CpG island between nucleotide positions 51 and 1197 of p439-a23.fP (Figure 5.3C; Appendix II). This is in agreement with the relatively high number of rare cutting restriction enzyme sites found within and flanking this region in the the PAC clone (Figure 5.3B&C). The size of the fusion transcript identified by Northern blot (Figure 4.1) predicted that the gene trap vector had

Figure 5.3: Characterisation of genomic DNA containing Group I and Group II sequences.

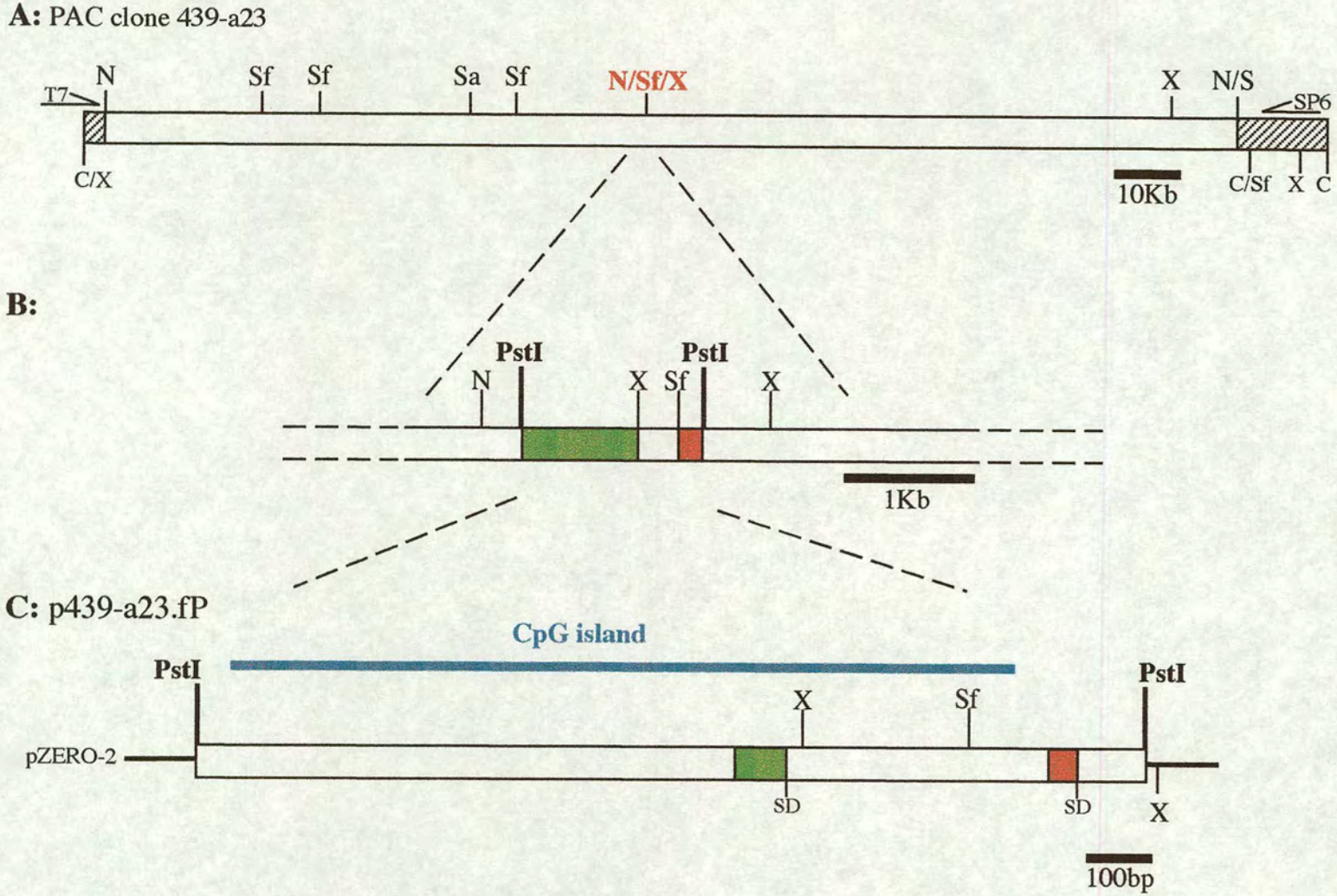
A: Restriction map of PAC clone 439-a23.
The genomic insert is approximately 190Kb. Additional XhoI sites crudely mapped between the T7 end and the NotI/SfiI/XhoI sites. The Group I (L-69) and Group II (UBT-1) fusion sequences map to the region of the NotI/SfiI/XhoI sites at the centre of the genomic insert (highlighted in red). Hatched areas correspond to the pPAC4 vector sequences.
Scale: 1cm to 10Kb.

B: A 1.5Kb PstI genomic fragment contains the Group I and Group II sequences.
Higher resolution restriction map of the genomic region surrounding the NotI/SfiI/XhoI sites. The Group I sequence maps to the 1Kb PstI/XhoI fragment and the Group II sequence to the 200bp SfiI/PstI fragment.
Scale: 2cm to 1Kb.

C: Subclone p439-a23.fP
Sequencing of the 1.5Kb PstI fragment enabled the precise mapping of the Group I and Group II fusion sequences. Both splice donors (SD) used by the gene trap vector are in the same orientation within this sequence. The GRAIL CpG island algorithm identified the sequence between nucleotides 51 and 1197 as being an excellent CpG island (blue bar).
Scale: 1cm to 100bp.

C, ClaI; Sa, SalI; Sf, SfiI; N, NotI; X, XhoI.

Figure 5.3:


A: PAC clone 439-a23
T7
N
Sf
Sf
Sa
Sf
N/Sf/X
X
N/S
SP6
C/X
10Kb
C/Sf
X
C
B:
PstI
PstI
N
X
Sf
X
1Kb
C: p439-a23.fP
CpG island
PstI
PstI
X
Sf
pZERO-2
SD
SD
X
100bp

integrated into the 5' end of the endogenous transcript. The identification of a CpG island in this region supports this prediction.

(b) Group I and Group II fusion sequences

The sequence analysis of the ubiquitously expressed Group I fusion transcript predicted that translation of *lacZ* would be out-of-frame and no β-gal acitvity would be produced from this transcript. The Group II fusion transcript expression correlated with I114 reporter activity and was subsequently predicted to produce reporter activity.

Analysis of the nucleotide sequence of the splice donor site used by the gene trap vector in splicing to the ubiquitous Group I fusion sequence reveals a good overall match to the splice donor consensus (Alberts *et al.*, 1994; Figure 5.4A). However, a cytosine residue replaces the almost invariant thymine at position 905 in the intronic sequence (Figure 5.4A). This may explain why none of the exon prediction algorithms available on the NIX program identified the splice donor site used by the gene trap vector at position 903 (Appendix II). Although the different exon prediction algorithms identified several different splice donor sequences, the majority of these programs identified the same open reading frame from a translational start codon at position 523 (Appendix I). This ORF corresponds to translation of the Group I fusion sequence in frame 2 in the context of the gene trap integration (Figure 5.4A; Figure 4.2). The isolation and conceptual translation of the corresponding cDNA sequence of this genomic region confirms position 903 as the endogenous splice donor and frame 2 of the Group I sequence as the ORF of the endogenous gene (see Chapter 6). Therefore, as predicted in Chapter 4, *lacZ* will not be translated from the Group I fusion transcript and no β-gal activity will be produced (Figure 5.5A).

An excellent correlation is observed betweeen the splice donor sequence used by the gene trap vector in splicing to the Group II liver specific fusion sequence and the consensus splice donor sequence (Figure 5.4B). Using NIX analysis, two exon prediction programs (Fex and FGene) identified the same splice donor site at position

Figure 5.4: Analysis of 439-a23.fP genomic sequence

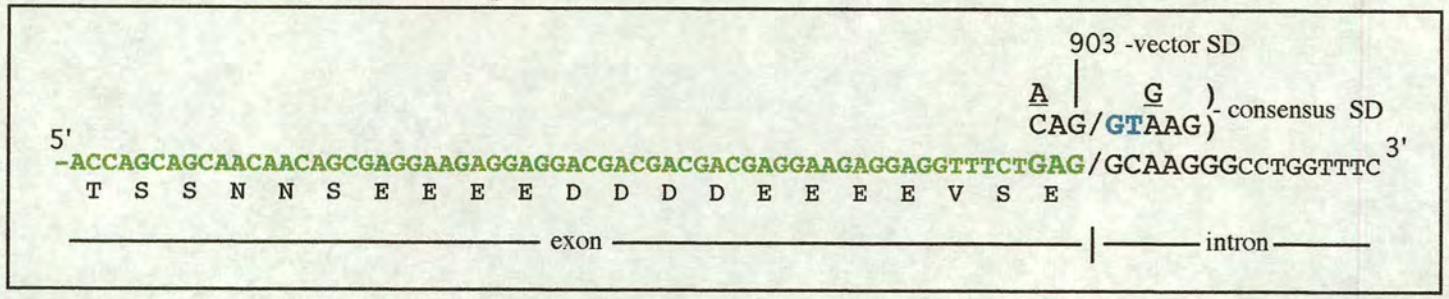
A: Genomic sequence of the Group I splice donor site.
The Group I fusion sequence (highlighted in green) from p439-a23.fP is compared to the splice donor consensus sequence. Position 903 corresponds to the splice donor site used by the gene trap vector. The nearly invariant G and T residues within the splice donor consensus are highlighted in blue. The Group I fusion sequence has been translated in-frame with *lacZ* (frame 2 - Figure 4.2).

B: Group II genomic sequence in context of NIX predicted exon
The Group II fusion sequence is highlighted in red. Comparison of the splice donor site used by the gene trap vector (position 1379) and the NIX identified splice acceptor sequence at position 1301 is compared to the consensus splice donor and acceptor sequences respectively. Nearly invariant residues within these consensus are highligted in blue. Translation of this predicted exon identifies an ORF corresponding to frame 1 in the context of the Group II fusion transcript (Figure 4.2).
(i) Splicing of the exon containing the Group II sequence (via SA at 1301) to the upstream Group I sequence (via SD at 903) maintains the predicted ORF.

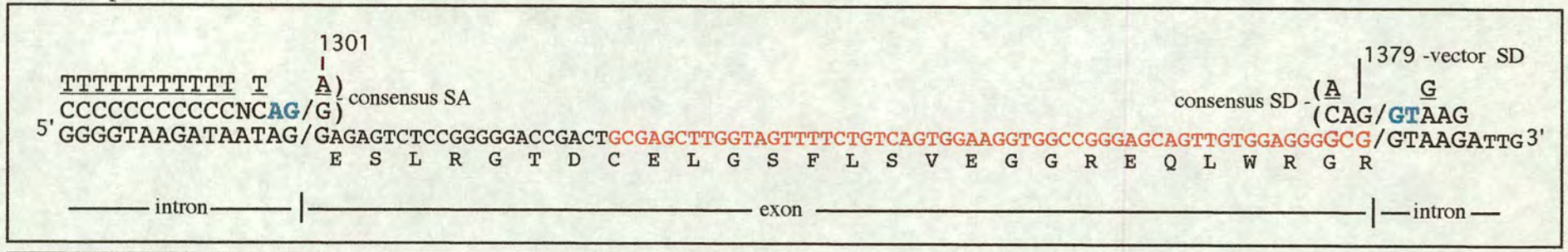
C: Group II genomic sequence in context of NIX predicted promoter
Predicted promoter sequence immedately upstream of the Group II fusion sequence. Transcriptonal start site is at position 1320 with the TATA box identified at 1292 (promoter position -28).

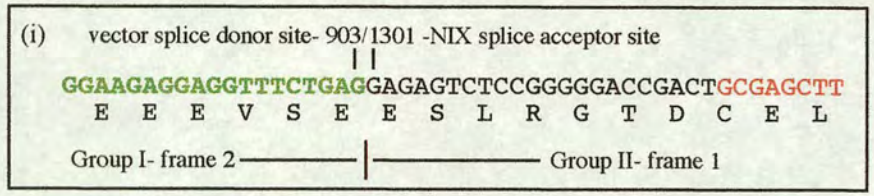
Figure 5.4:

A: Group I genomic splice donor sequence

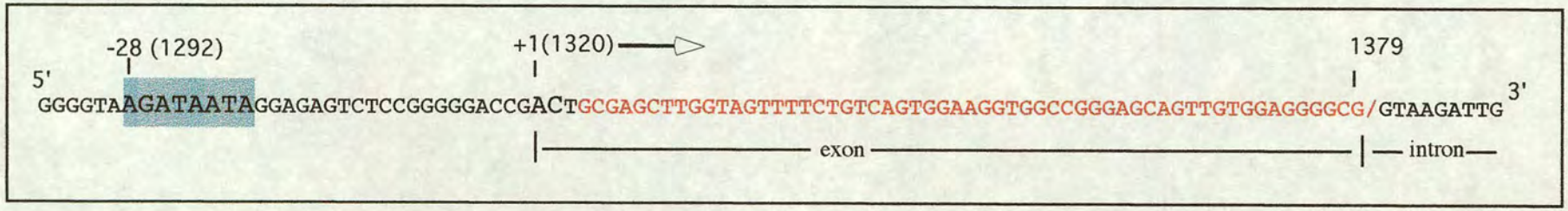


B: Group II exon?





C: Group II promoter?



1379 (Appendix II). However, the NIX analysis of the sequence immediately upstream of the Group II fusion transcript identifies this sequence as either an internal exon or the first exon downstream of a RNA polymerase II promoter.

The two exon prediction programs which identified the Group II splice donor sequence at position 1379 also predicted an upstream splice acceptor site to be at position 1301 (Figure 5.4B). It is of interest to note, however, that this putative splice acceptor does not have a consensus polypyrimidine stretch upstream (Figure 5.4B). Conceptual translation of this predicted exon produces 26 amino acid residues with no stop codons, corresponding to frame 1 of the Group II fusion transcript. In the context of gene trap vector splicing, this would result in the in-frame translation of the *lacZ* gene generating β-gal activity from this transcript. Splicing of this exon to the Group I splice donor site used by the gene trap vector at position 903 (Figure 5.4B(i)) maintains the ORF corresponding to frame 2 for the Group I sequence (endogenous cDNA ORF) and frame 1 for the Group II sequence (Figure 4.2 & 5.4B).

The transcriptional start site Wingender (TSSW) algorithm predicts human RNA polymerase II promoter regions and transcriptional start sites. The predictions are based on the presence of transcription factor binding sites coupled with the oligonucleotide composition of predicted transcriptional start sites. The TSSW algorithm predicted an excellent promoter with a TATA box 28 nucleotides upstream of the predicted transcriptional start site at position 1320 (Figure 5.4C). However, although this region is AT rich, it does not fit the TATA box consensus of TATAa/tAa/t (Breathnach and Chambon, 1981). Additional general transcription factor binding sites were identified upstream including a GC box which is bound by Sp1 (Appendix I). The transcriptional start site is immediately upstream of the Group II liver specific fusion transcript sequence (Figure 5.4C). No start codon is present within this potentially transcribed sequence from position 1320. Therefore, the production of β-gal activity from this sequence, when spliced to the gene trap vector, would require the translation of *lacZ* from its own start codon.

5.5. Summary

Although there are inherent inaccuracies in predicting sequence function from computer algorithms, the results of the NIX analysis identifies two potential mechanisms which could be responsible for the liver specific expression of the Group II fusion transcript. The Group II fusion sequence may represent an internal exon of the endogenous trapped gene. The expression of the Group I fusion transcript suggests that the endogenous gene is ubiquitously expressed. Therefore, expression of the Group II exon in the endogenous gene may be under the control of a liver specific splicing mechanism. If the Group II sequence represents an internal exon, splicing of the gene trap vector results in the ORF of the endogenous gene being in-frame with *lacZ* which will result in the production of reporter activity (Figure 5.5Bi).

The other possibility is that the Group II fusion sequence represents an alternative first exon of the endogenous gene which is expressed from a liver specific promoter independently of the promoter driving expression of the ubiquitous Group I fusion sequence (Figure 5.5Bii). If the Group II sequence represents the first exon of the gene, then the lack of a methionine start codon identifies this sequence as being untranslated. Therefore, *lacZ* will be translated from its own translational start codon producing active β-gal (Figure 5.5Bii).

The evidence for and against these two possibilities and their consequences in the context of the full length cDNA are discussed at the end of Chapter 6.

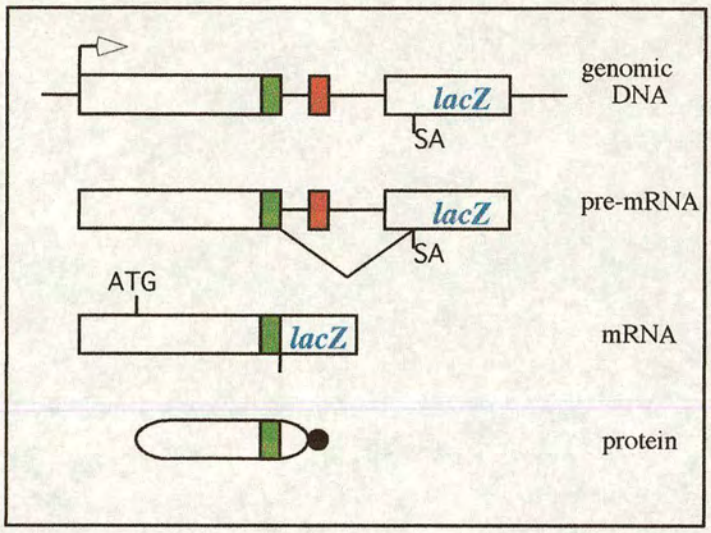
Figure 5.5: Summary of the I114 gene trap integration event

A: Splicing of the gene trap vector to the Group I fusion sequence
The Group I sequence is ubiquitously expressed from the endogenous promoter. The Group I fusion transcript is translated from the endogenous upstream start codon which places *lacZ* translation out-of-frame. No β-gal activity is produced from this fusion transcript.

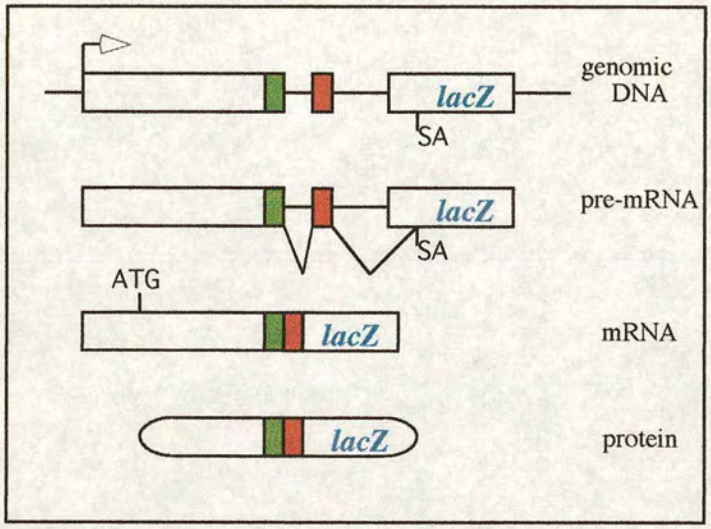
B: Gene trap vector splicing to the Group II fusion sequence
(i) Group II sequence as an internal exon
The Group I containing exon is ubiquitously expressed from the endogenous promoter. In the foetal liver, the Group II fusion sequence is alternatively spliced to the gene trap vector and the upstream Group I fusion sequence. The fusion transcript is translated from the endogenous upstream start codon which will place translation of *lacZ* in-frame. β-gal activity will be produced from this transcript.
(ii) Group II sequence expressed from an alternative promoter
In the foetal liver, an alternative promoter drives expression of the Group II fusion sequence. Splicing to the gene trap vector produces the Group II fusion transcript which will be translated from the *lacZ* start codon because the Group II sequence is untranslated. β-gal activity will be produced from this transcript.

Figure 5.5:

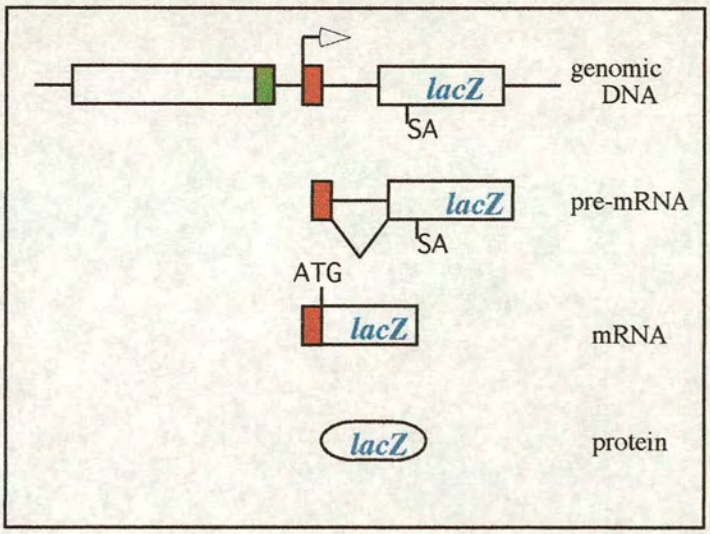
A:



B(i):



B(ii):



Chapter 6

RESULTS

Isolation and Characterisation of Gene Trap Ankyrin Repeat (*gtar*)

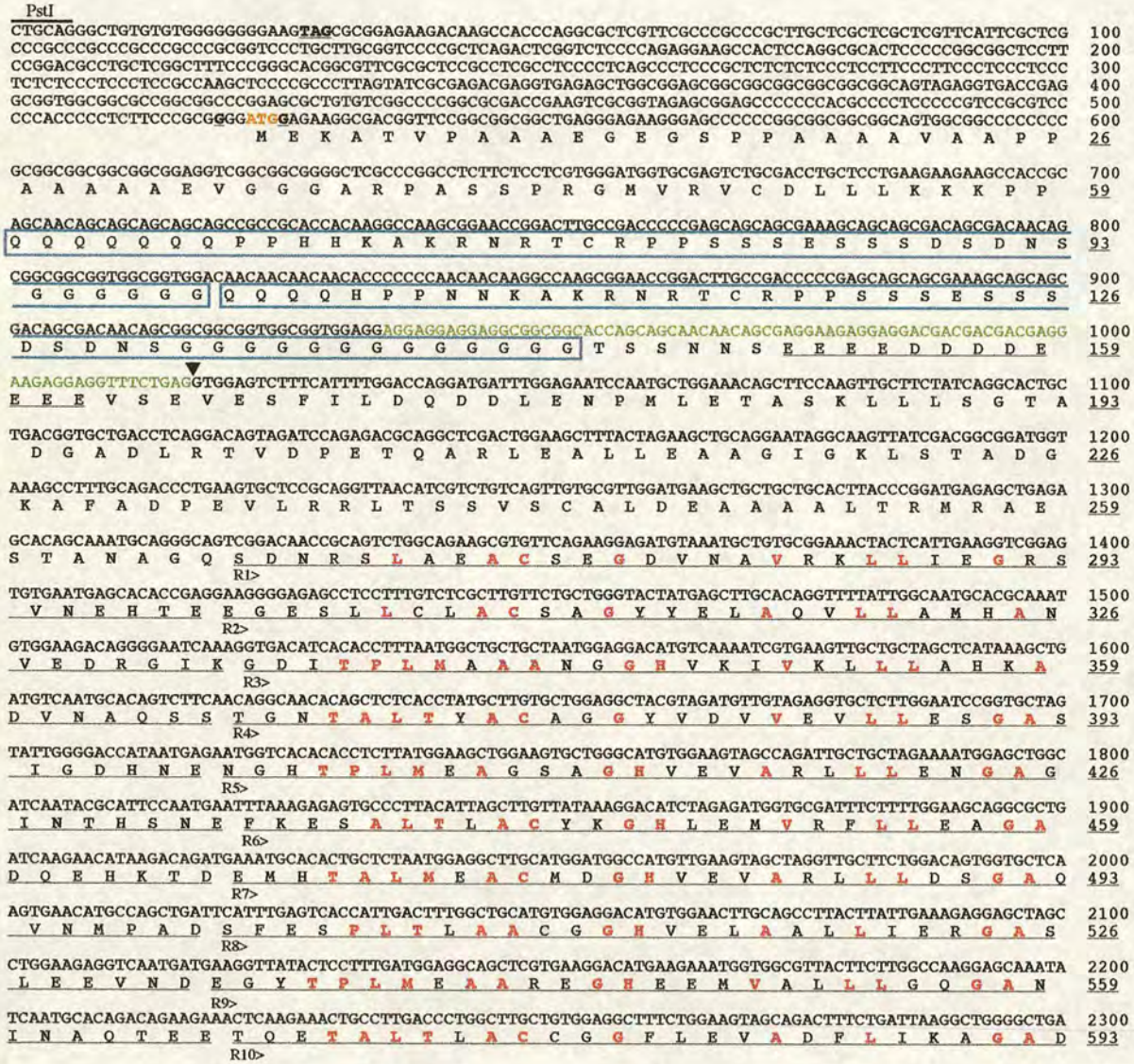
6.1. cDNA Library Screening Using the Group I Sequence

The Group I and Group II fusion sequences show a close association at the genomic level. In order to isolate the gene associated with this genomic region, cDNA library screening was carried out with the Group I fusion sequence. The fusion transcript L-A8 (Figure 4.2) was used to probe a random primed R1 ES cell cDNA library cloned into the λZAPII vector (gift from Dr Hitoshi Niwa). Fifty positive plaques were identified from the primary screening of $1x10^{-6}$ clones of which 4 clones (p1Bα-2, p3Aα-2, p4Aα-1 and p5Aα-1) were characterised and showed overlapping restriction digest patterns (Appendix III). The largest clone (p1Bα-2) is 4kb in size and spans the sequence contained in clones p3Aα-2, p4Aα-1 and p5Aα-1 (Appendix III). To isolate cDNA sequence downstream of these clones, a probe derived from the most 3' sequence of clone p1Bα-2 (Appendix III) was used to screen the same R1 ES cell cDNA library. A further 3 clones were isolated (pGC3, pGC7 and pGC10) which overlapped with the 3' end of clone p1Bα-2 (Appendix III).

The sequencing strategy outlined in Appendix III produced a contiguous sequence of 6139 nucleotides. Conceptual translation of this sequence identifies an open reading frame encoding a protein of 1637 amino acids with a predicted molecular weight of 174 KD. The translational stop codon at nucleotide position 28 identifies the first in-frame methionine translational start codon as being at position 523 (Figure 6.1). Moreover, this putative start codon is in the context of a strong Kozak consensus sequence for translational initiation (Kozak, 1996, Figure 6.1). We have named this

Figure 6.1: Nucleotide and protein sequence of *gtar*

- The PstI site identifies the same 5' limit of sequence derived from cDNA and PAC fragment
- Start codon: **ATG** at position 523 with bold underlined nucleotides at 520(-3) and 526(+4) in agreement with the Kozak consensus (RNNatgG, where R=purine). **TAG** at position 28, in-frame stop.
- Novel tandem repeat sequence flanked by Q and G residues boxed in **blue**.
- Group I fusion sequence is highlighted in **green** with potential gene trap vector integration site marked by arrow.
- Hyperacidic cluster underlined.
- Ankyrin repeats underlined and numbered (R1-R25) with highly conserved residues in **red**.
- Splice sites defining the different splice variants are arrowed at positions 2957, 3707, 3836 and 5058.
- NLS residues highlighted in **orange** with **orange** undelining showing the 3 different NLS units.



gene *gtar* (**g**ene **t**rap **a**nkyrin **r**epeat). The presence of translational stop codons in all three reading frames at the 5' and 3' ends of the contiguous sequence identifies the 5' and 3' UTRs, which suggests that the 6139bp contiguous sequence contains the complete coding sequence of *gtar*. However, no poly A addition signal (AATAAA) has been identified in the 3' UTR, suggesting sequence is missing from the 3' end of this cDNA sequence. None of the cDNAs isolated contained Group II fusion sequences as determined by the Southern blotting of the seven cDNA clones with the L-69 RACE clone.

6.2. Analysis of the Group I Sequence

The Group I fusion sequence is within the coding region of the cDNA, between position 938 and 1017 (Figure 6.1), suggesting insertion of the gene trap vector has occurred into an intron between the splice donor site at position 1017 and the putative downstream endogenous splice acceptor site at position 1018. Furthermore, it indicates that the same splice acceptor site used by the gene trap vector is used in the context of *gtar* splicing. The cDNA sequence from position 1 to position 1017 matches exactly the sequence from position 1 to 1017 of the genomic sequence derived from the 439-a23 PAC clone (p439-a23.fP). This identifies nucleotides 1 to 1017 of *gtar* as a single exon at the genomic level. The ORF identified from *gtar* corresponds to translation of the Group I fusion transcript in frame 2 (Figure 4.2) confirming that the Group I fusion transcript is translated from the endogenous upstream start codon resulting in the out-of-frame translation of the *lacZ* gene. β-gal activity is not therefore generated from this gene trap splicing event.

6.3. Protein Sequence Analysis of *gtar*

(a) Ankyrin Repeats

The BLASTP algorithm was used to compare *gtar* protein sequence against a non-redundant GenBank protein database. BLASTP identified significant homology between the protein sequence and a number of ankyrin-like (ANK) repeat containing genes (data not shown). This homology lead to the identification of a total of 25 ANK repeats ranging from 30-35 amino acids in length (Figure 6.2). The consensus from these repeats matches well with the human erythrocyte ankyrin repeat consensus (Lux *et al.*, 1990; Figure 6.2). The ANK repeats are organised into two separate domains of 15 and 10 repeats (Figures 6.1 & 6.2). The terminal ANK repeats of these domains are more divergent from the ANK consensus sequence than the centrally located repeats, a common feature of ANK repeat domains (Bork, 1993).

(b) Hyperacidic Cluster

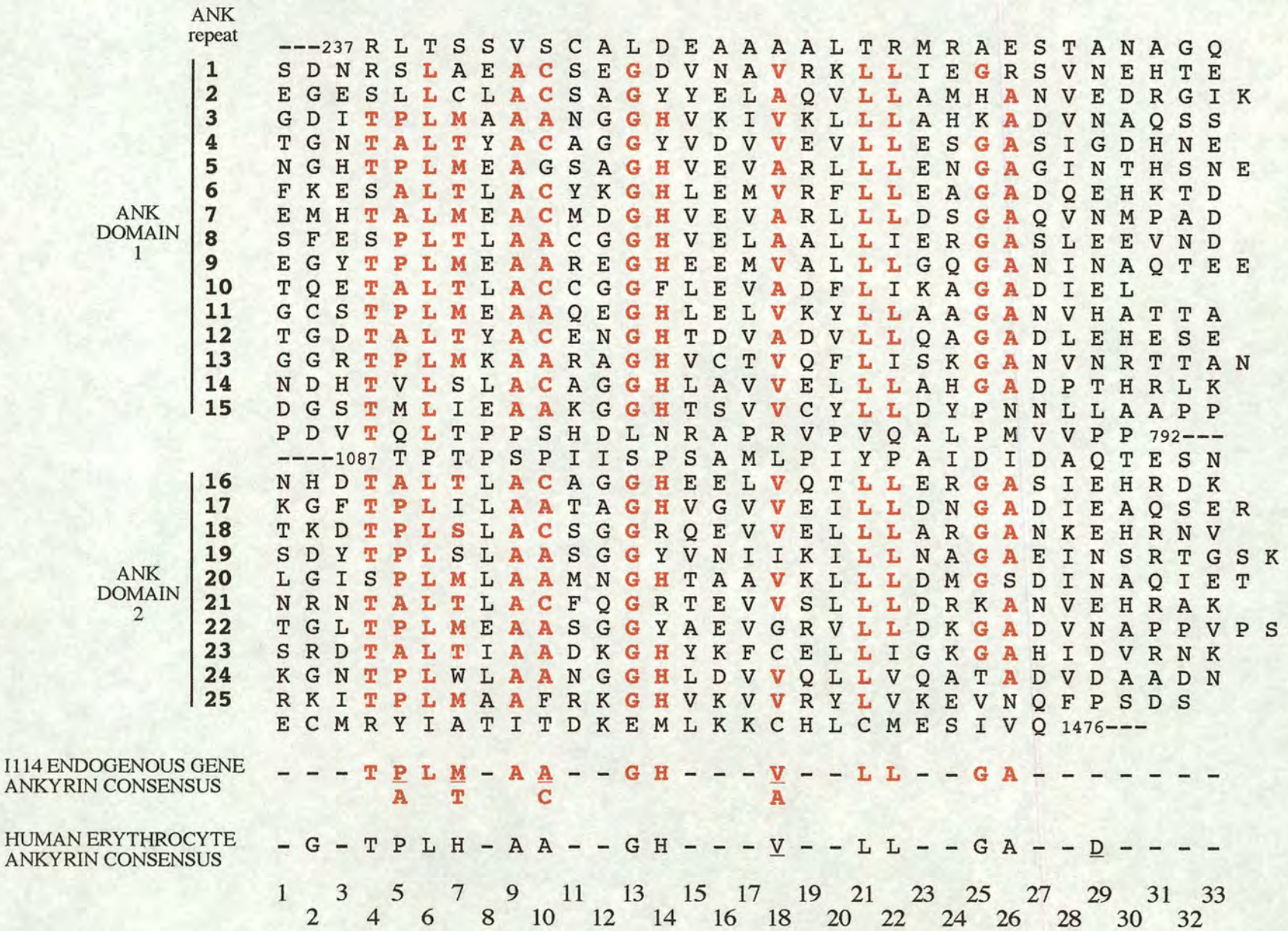
Such strong homology to ANK repeat containing genes may obscure protein homologies outwith the ANK repeat domains. Therefore, the N-terminal amino acid residues 1 to 266, residues 769 to 1114 between the ANK repeat domains 1 and 2 and and the C-terminal residues 1450 to 1637 were compared to the GenBank protein database using BLASTP.

The 13 consecutive aspartate and glutamate residues between amino acid position 151 and 162 (within the Group I fusion sequence, Figure 6.1) represent a region with a strong local negative charge. Such hyperacidic clusters were identified in a number of different transcription factors including the winged helix transcription factors human *FREAC-4* (accession nos. U59832), CWH-1 (chick, U37272) and MBF-2 (mouse, L38607), the zinc finger transcription factor δ/UCRBP (mouse, M74590) and the nucleolar transcription factor UBF-1 (mouse, P25976). Studies on

Figure 6.2: Optimal alignment of the *gtar* ANK repeats

The alignment of the amino acid sequence of the I114 endogenous gene from residues 237-792 and 1087-1476 identifies a total of 25 tandem ANK repeats arranged into two domains of 15 repeats (Domain 1) and 10 repeats (Domain 2).
Amino acids conserved in at least two thirds of the repeats are are highlighted in red and summarised in the consensus below. A good corrrelation is seen between the I114 endogenous gene ANK consensus and the human erythrocyte ANK consensus (Lux *et al.*, 1990). ANK repeat residues are numbered 1-33.

Figure 6.2:


ANK repeat
ANK DOMAIN 1
ANK DOMAIN 2
---237 R L T S S V S C A L D E A A A A L T R M R A E S T A N A G Q
1 S D N R S L A E A C S E G D V N A V R K L L I E G R S V N E H T E
2 E G E S L L C L A C S A G Y Y E L A Q V L L A M H A N V E D R G I K
3 G D I T P L M A A A N G G H V K I V K L L L A H K A D V N A Q S S
4 T G N T A L T Y A C A G G Y V D V V E V L L E S G A S I G D H N E
5 N G H T P L M E A G S A G H V E V A R L L L E N G A G I N T H S N E
6 F K E S A L T L A C Y K G H L E M V R F L L E A G A D Q E H K T D
7 E M H T A L M E A C M D G H V E V A R L L L D S G A Q V N M P A D
8 S F E S P L T L A A C G G H V E L A A L L I E R G A S L E E V N D
9 E G Y T P L M E A A R E G H E E M V A L L L G Q G A N I N A Q T E E
10 T Q E T A L T L A C C G G F L E V A D F L I K A G A D I E L
11 G C S T P L M E A A Q E G H L E L V K Y L L A A G A N V H A T T A
12 T G D T A L T Y A C E N G H T D V A D V L L Q A G A D L E H E S E
13 G G R T P L M K A A R A G H V C T V Q F L I S K G A N V N R T T A N
14 N D H T V L S L A C A G G H L A V V E L L L A H G A D P T H R L K
15 D G S T M L I E A A K G G H T S V V C Y L L D Y P N N L L A A P P
P D V T Q L T P P S H D L N R A P R V P V Q A L P M V V P P 792---
----1087 T P T P S P I I S P S A M L P I Y P A I D I D A Q T E S N
16 N H D T A L T L A C A G G H E E L V Q T L L E R G A S I E H R D K
17 K G F T P L I L A A T A G H V G V V E I L L D N G A D I E A Q S E R
18 T K D T P L S L A C S G G R Q E V V E L L L A R G A N K E H R N V
19 S D Y T P L S L A A S G G Y V N I I K I L L N A G A E I N S R T G S K
20 L G I S P L M L A A M N G H T A A V K L L L D M G S D I N A Q I E T
21 N R N T A L T L A C F Q G R T E V V S L L L D R K A N V E H R A K
22 T G L T P L M E A A S G G Y A E V G R V L L D K G A D V N A P P V P S
23 S R D T A L T I A A D K G H Y K F C E L L I G K G A H I D V R N K
24 K G N T P L W L A A N G G H L D V V Q L L V Q A T A D V D A A D N
25 R K I T P L M A A F R K G H V K V V R Y L V K E V N Q F P S D S
E C M R Y I A T I T D K E M L K K C H L C M E S I V Q 1476---
I114 ENDOGENOUS GENE ANKYRIN CONSENSUS
- - - T P L M - A A - - G H - - - V - - L L - - G A - - - - - - - -
A T C A
HUMAN ERYTHROCYTE ANKYRIN CONSENSUS
- G - T P L H - A A - - G H - - - V - - L L - - G A - - D - - - -
1 3 5 7 9 11 13 15 17 19 21 23 25 27 29 31 33
2 4 6 8 10 12 14 16 18 20 22 24 26 28 30 32

UBF-1 have identified the hyperacidic cluster as being essential for transcriptional activation (Voit *et al.*, 1992). As well as transcription factors, hyperacidic clusters are present in functionally diverse proteins for example β-tubulin (rice, X79367) and the neuronal anion exchanger AE3 (mouse, M28383). None of the other protein regions entered using BLASTP showed a significant level of homology to database sequences.

(c) Bipartite Nuclear Localisation Signals

Computer-based motif searching was performed on the full protein sequence using the suite of protein analysis programs available at the HGMP Resource Centre (http://www.hgmp.mrc.ac.uk). The PSort program searches for motifs involved in determining the sub-cellular localisation of the protein. The amino acid sequences from residues 1506 to 1522, 1516 to 1533 and 1590 to 1607 correspond to the consensus sequence for the bipartite nuclear localisation signal (NLS) initially identified from the nucleoplasmin protein of *Xenopus laevis* (Figure 6.1). The bipartite NLS consists of two clusters of basic amino acids separated by a 10 amino acid spacer and has been found in more than 50% of nuclear proteins, but only 4% of the non-nuclear proteins analysed from a sequence database (Dingwall and Laskey, 1998; Hicks and Raikel, 1995). The same three putative bipartite NLS were also identified using the ProfileScan program which searches protein profile databases for protein motifs (http://expasy. hcuge.ch/). Moreover, within the region containing the bipartite NLS, a simpler NLS consensus identified from the SV40 large T antigen protein consisting of four consecutive basic residues is also present (Hicks and Raikhel, 1995). The whole region containing the three putative NLS (1506-1607) is rich in basic amino acid residues (Figure 6.1).

(d) Identification of a Novel Tandem Repeat

Another intriguing feature of the protein sequence is the presence of two highly conserved tandem repeat sequences at the N-terminus. Comparison of the nucleotide sequence between nucleotide position 700 to 819 and position 820 and 954 identifies 92% homology between these sequences. This homology is preserved at the amino acid level with residues 62-99 aligning against residues 100-137 (Figure 6.3A). A single repeat unit consists of flanking homopolymeric runs of glutamine residues at the N-terminal end and glycine residues at the C-terminal end (Figure 6.3A). Within the flanking homopolymeric residues, a 28 amino acid motif is present with 25 out of 28 (89%) of the residues identical between the two repeats (Figure 6.3A). This core amino acid sequence is rich in proline, serine, glutamic acid, threonine and aspartic acid residues 16/28 (57%, Figure 6.3A). Regions high in these residues flanked by basic residues are a feature of PEST sequences characteristic of proteins with a short half-life (Rogers *et al.*, 1986). Although potential PEST sequences are not contained absolutely within each repeat, residues 79 to 104 and 117 to 186 are identified as PEST sequences (Figure 6.3B; Rogers *et al.*, 1986). It is of interest to note that the second putative PEST sequence (117-186) also incorporates the hyperacidic cluster. No significant level of homology was observed between the nucleotide sequence of both repeats (700-954) and sequences in both the the non-redundant GenBank database and the expressed sequence tag database dbEST using BLASTN. Similary, the protein sequence of both repeats between amino acid residues 60-144 showed no homology to protein sequence databases or dbEST translated in all six reading frames.

6.4. Isolation of *gtar* Splice Variants

The full *gtar* cDNA sequence presented above consists essentially of the 5' and 3' ends of clones p1Bα-2 and pGC7 respectively to give the full coding region. However, sequences present in the full cDNA presented in Figure 6.1 are absent in

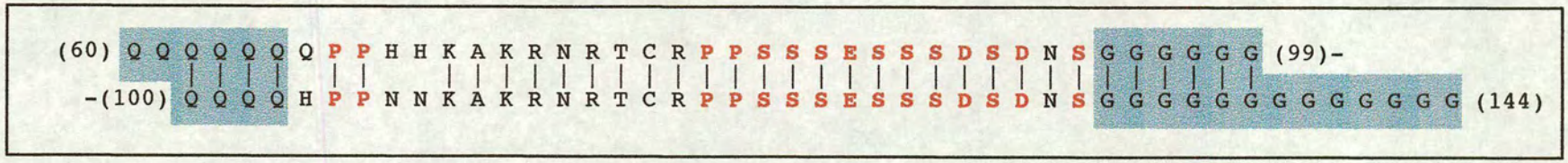
Figure 6.3: Identification of a novel tandem repeat sequence

A: Optimal alignment of the amino acid residues 60 to 99 with 100 to 144 identifies a core repeat sequence showing 89% identity between the two repeats flanked by homopolymeric runs of glutamine residues at the N-terminal end and glycine residues at the C-terminal end (boxed in blue). The residues highlighted in red correspond to amino acid residues associated with PEST sequences.

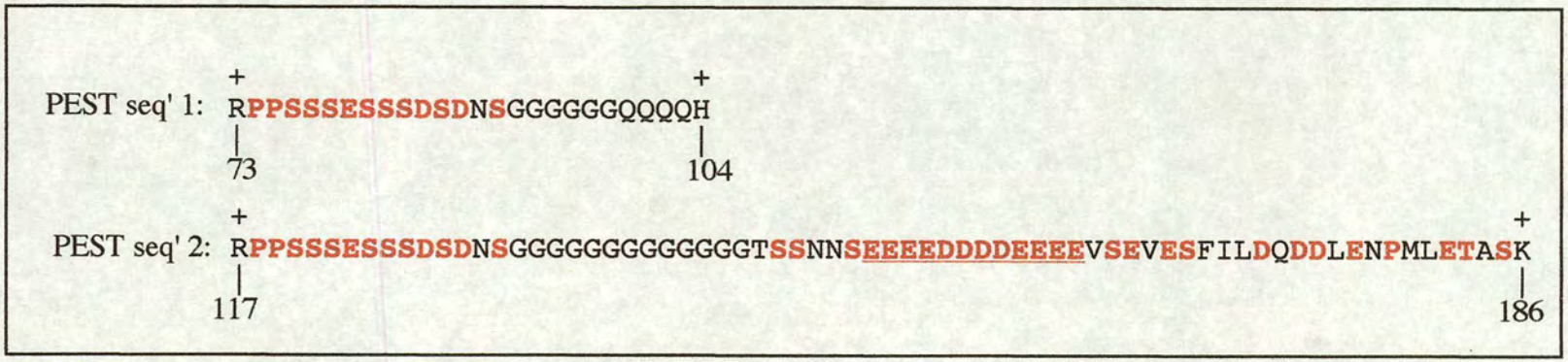
B: Amino acid sequence of two potential PEST sequences.
The PEST sequence consists of P, E, S, T, and to a lesser extent D residues (highlighted in red) flanked by a least a single positive amino acid residue (+). PEST sequence 1 is from residue 74- 104 and is contained absolutely within the two novel repeat sequences. PEST sequence 2 between residues 117 and 186 spans the second novel repeat sequence and downstream sequences including the hyperacidic cluster between residues 151 and 162 (underlined).

Figure 6.3:

A:


(60) Q Q Q Q Q Q Q P P H H K A K R N R T C R P P S S S E S S S D S D N S G G G G G G (99)-
-(100) Q Q Q Q H P P N N K A K R N R T C R P P S S S E S S S D S D N S G G G G G G G G G G G G (144)

B:


PEST seq' 1: RPPSSSESSSDSDNSGGGGGGQQQQH
73
104
PEST seq' 2: RPPSSSESSSDSDNSGGGGGGGGGGGGGTSSNNSEEEEDDDDEEEEVSEVESFILDQDDLENPMLETASK
117
186

clones p1Bα-2, pGC3 and pGC10 suggesting that these clones represent *gtar* splice variants (Figure 6.4). The nucleotide sequence between position 2957 and 3707 is absent in clones p1Bα-2, pGC3 and pGC10 (Figure 6.4). The ORF of the full length cDNA is maintained but amino acid residues between position 812 to 1063 are eliminated. The putative protein sequence of this eliminated sequence showed no homology to existing protein sequences or motifs in the database.

The sequence of clones p1Bα-2 and GC3 continues into the ANK repeat domain 2 and the sequence of these clones terminates at position 3904 and 4680 respectively. This sequence correponds to splice variant 1 (SV-1; Figure 6.4). Assuming that the 5' and 3' ends of SV-1 correspond to the same sequence as the full length cDNA (Figure 6.4), then SV-1 would represent a 5389bp transcript producing a protein of 1387 amino acids (146kD).

Clone pGC10 sequence represents a different splice isoform (SV-2). Splicing between nucleotide position 3836 and 5058 maintains the ORF of the full length cDNA eliminating amino acid residues 1106 to 1513 from the protein sequence. This results in the deletion of ANK repeat domain 2 but maintains the putative NLS at position 1516 which (Figure 6.4) potentially represents a transcript of 4168bp producing a protein of 980 amino acids (103kD).

6.5. Genomic structure of *gtar*

In order to define the genomic structure of PAC clone 439-a23 relative to *gtar*, various regions of the cDNA were used to probe 439-a23 digested with rare cutting restriction enzymes. As previously identified, the oligonucleotide UBT-1 derived from the Group I fusion sequence (corresponding to nucleotide residues 966-1007 from the cDNA) is located towards the centre of the genomic insert around the NotI/SfiI/XhoI restriction sites (Figures 6.5).

The probe GC10.fXb is derived from clone pGC10 and corresponds to nucleotides 1528-1870 (Appendix III). This cDNA fragment hybridises to the 80kb

Figure 6.4: Splice variants of *gtar*

The total cDNA sequence is represented at the top of the diagram. The open boxed region corresponds to translated sequence with the solid black lines representing untranslated regions. The nucleotide position of each clone relative to the full length cDNA is given above and the corresponding amino acid residue below each clone.
ANK repeats are indicated by the numbered blue boxes.
Yellow boxes denote the novel tandem repeat.
The green box shows the hyperacidic cluster.
Orange boxes denote the bipartite NLS.

Scale 1.5cm to 500bp

Figure 6.4:

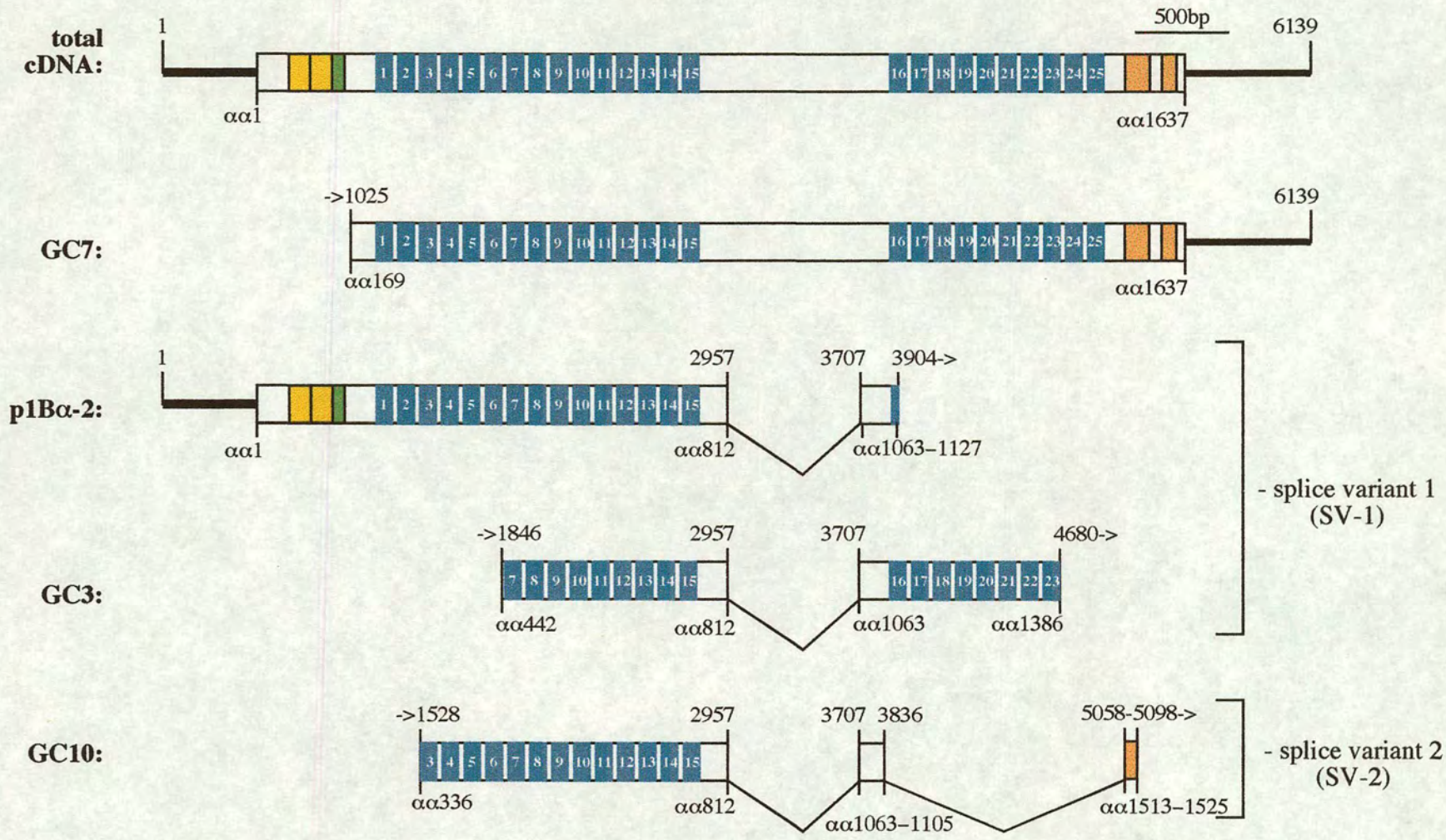
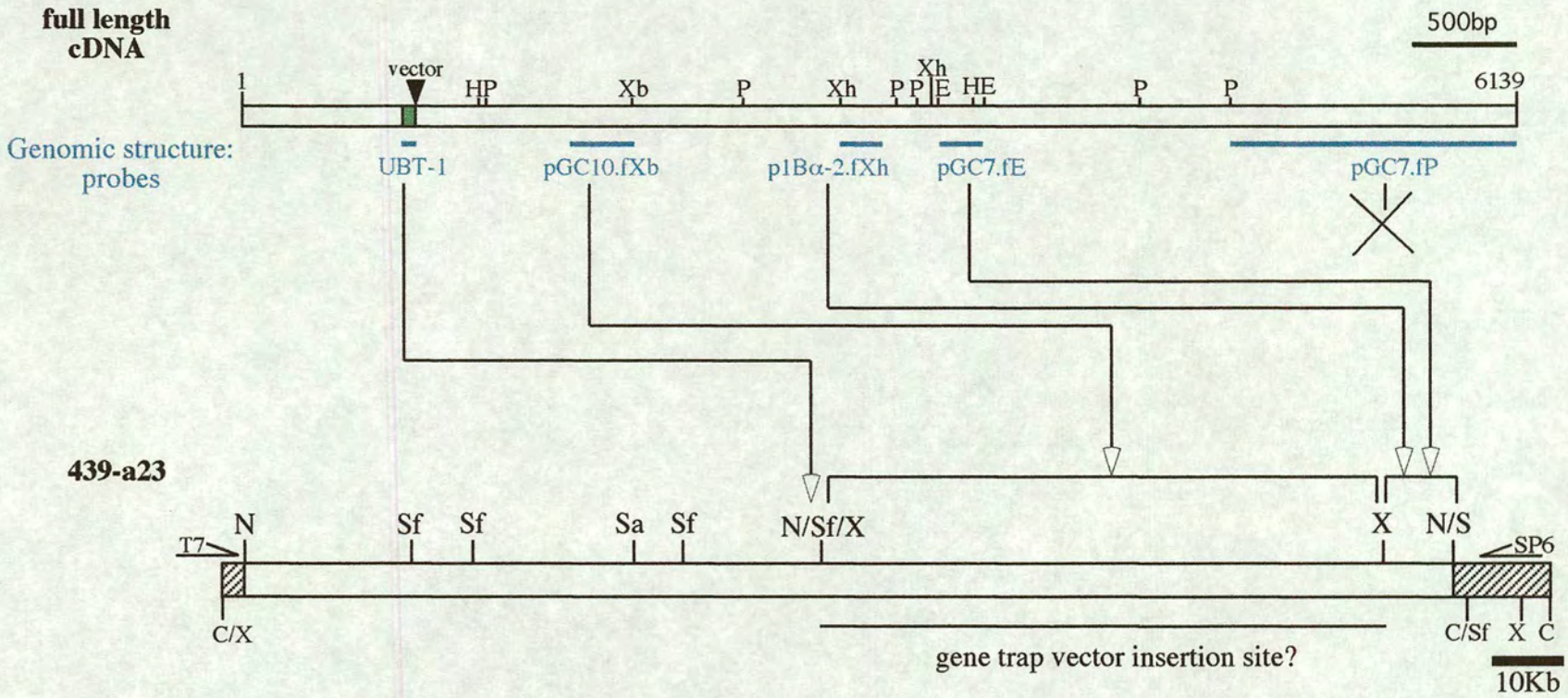

total cDNA:
1
500bp
6139
αα1
αα1637
GC7:
->1025
6139
αα169
αα1637
p1Bα-2:
1
2957
3707
3904->
αα1
αα812
αα1063–1127
GC3:
->1846
2957
3707
4680->
αα442
αα812
αα1063
αα1386
- splice variant 1 (SV-1)
GC10:
->1528
2957
3707
3836
5058-5098->
αα336
αα812
αα1063–1105
αα1513–1525
- splice variant 2 (SV-2)

Figure 6.5: Genomic structure of *gtar*

Diagram showing the where different probes derived from the full length endogenous cDNA mapped to the restriction map of the genomic DNA from PAC clone 439-a23. Restriction map of the full length cDNA clone shows the Group I sequence (green box) and the potential gene trap vector integration site (arrowed). The position and names of the cDNA fragments used to probe the PAC clone 439-a23 are represented as blue bars underneath the cDNA. The arrows indicate which cDNA probe mapped to which restriction fragment from the PAC clone genomic DNA. Fragment pGC7.fP failed to hybridise to 439-a23 (crossed). The predicted genomic gene trap vector insertion site is underline

Scale 1.5cm to 500bp for cDNA, 1cm to 10kb for 439-a23.
H, HindIII; P, PstI; Xb, XbaI; Xh, XhoI; E, EcoRI; Sf, SfiI; N, NotI; Sa, SalI; C, ClaI.

Figure 6.5:


full length cDNA
500bp
vector
1
HP
Xb
P
Xh
P P
Xh
E
HE
P
P
6139
Genomic structure: probes
UBT-1
pGC10.fXb
p1Bα-2.fXh
pGC7.fE
pGC7.fP
439-a23
T7
N
Sf
Sf
Sa
Sf
N/Sf/X
X
N/S
SP6
C/X
C/Sf
X
C
gene trap vector insertion site?
10Kb

genomic DNA fragment towards the SP6 end of the genomic insert which contains none of the rare cutting restriction enzyme used, making it impossible to define more accurately its position within the genomic insert (Figure 6.5). Probe GC7.fE, corresponding to nucleotides 3310-3360 of the full length cDNA (Appendix III), hybridised to the 10kb XhoI genomic fragment closest to the SP6 end of the genomic insert. This same XhoI genomic fragment also hybridised to probe 1Bα2.fXh (nucleotides 2860-2957/3707-3904) derived from the SV-1 clone p1Bα–2 (Figure 6.5). Finally, probe pGC7.fP from the 3' end of the full length cDNA (nucleotides 4750-6139) did not hybridise to PAC clone 439-a23 (Figure 6.5).

The hybridisation pattern of probes UBT-1 and GC10.fXb which flank the predicted gene trap vector insertion site in the cDNA, suggests that the genomic integration site of the gene trap vector resides within the 80kb XhoI genomic fragment of the PAC clone 439-a23 (Figure 6.5). Probes UBT-1 and 1Bα2.fXh, which represent sequences separated by only 1800bp at the cDNA level, are approximately 80kb apart at the genomic level. This indicates that the genomic structure of the 5' end of *gtar* into which the gene trap vector has integrated consists of a single or several large intronic regions. Integration of the gene trap vectors into an intron distant from the Group I and Group II sequences would explain the failure to detect an RFLP between wild type and I114 genomic DNA (Appendix III).

6.6. Expression Analysis of *gtar*

(a) RNase Protection Assay Analysis

The results from the RPA performed in Chapter 4 using the Group I fusion transcript highlighted two important characteristics of the endogenous trapped gene in the context of the gene trap integration event. Firstly, the expression of the endogenous gene is more widespread than the liver specific reporter activity associated with the I114 gene trap integration. Secondly, the endogenous gene sequence was expressed

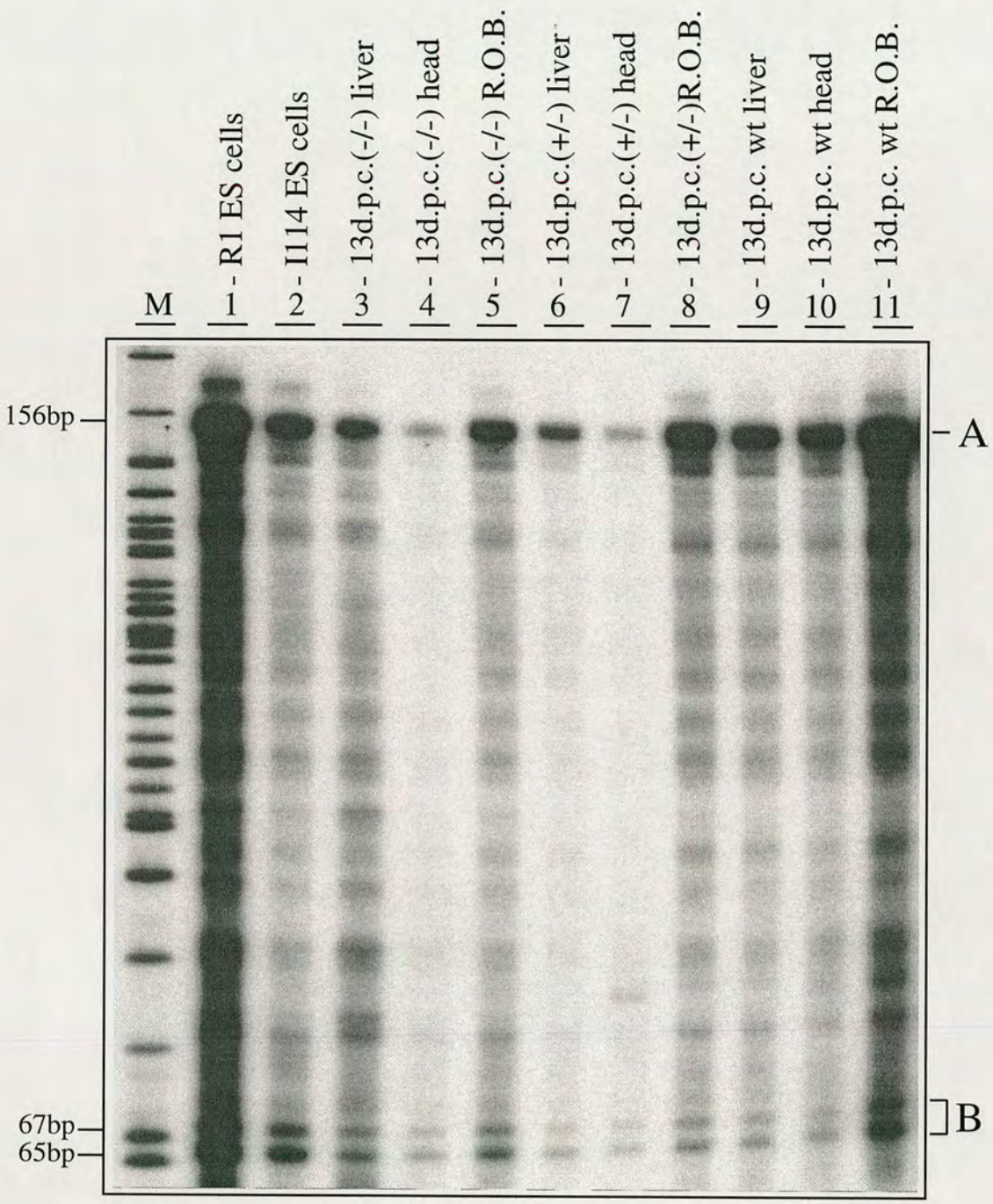
independently of the gene trap vector in I114 homozygous tissues indicating that the endogenous gene is splicing around the gene trap vector integration and producing wild type transcript (Figure 4.4).

An RPA assay was performed to investigate the expression pattern of *gtar* and also to assess the expression of the *gtar* sequences 3' of the putative gene trap integration site in I114 homozygous tissues. A riboprobe antisense to the endogenous cDNA sequence was generated from plasmid p5Aα.fPE(150). This plasmid contains the 156bp fragment from the 3' end of clone p5Aα-1 (Appendix III) which corresponds to nucleotides 2377 to 2533 of the full length cDNA clone (Figure 6.1). This riboprobe was hybridised to RNA isolated from R1 and I114 ES cells and liver, head and rest of body (R.O.B.) from wild type, I114 heterozygous and homozygous tissues. The RNA was digested and the products resolved by polyacrylamide gel electrophoresis. GAPdH was used as a loading control in all samples. The riboprobe is predicted to be protected as a 156bp fragment in tissues expressing this region of the endogenous cDNA. Figure 6.6 shows that the endogenous gene is expressed in all of the RNA samples analysed (lanes 1-11). The ubiquitous expression pattern of the I114 endogenous gene at mid-gestation is in agreement with the result of the RPA using the Group I fusion transcript. The expression of the endogenous gene sequence in I114 homozygous tissues (lane 3-5) confirms that the insertion of the gene trap vector has not generated a null allele. Wild type transcript is produced, presumably due to splicing around gene trap vector by *gtar* (Voss *et al.*, 1998b). However, the uneven loading of the RNA samples makes it difficult to determine if there is a reduced level of endogenous gene expression as a result of gene trap vector insertion. There is a possibility that this riboprobe, which includes part of *gtar* ANK repeat 11, all of repeat 12 and part of repeat 13 (Figure 6.1) is hybridising to other ANK repeat containing genes, resulting in the ubiquitous expression pattern. Although this specific probe has never been hybridised to genomic DNA, all the probes used in the RFLP analysis (Appendix III) which span other ANK repeats identify single copy DNA. This suggests that the *gtar* ANK repeats do not cross-hybridise to other ANK repeat containing genes.

Figure 6.6: Expression analysis of *gtar* by RNase protection

A riboprobe antisense to the cDNA sequence between positions 2377 and 2533 was hybridised to RNA from R1 and I114 ES cells, wild type, I114 homozygous and heterozygous liver, head and rest of body (R.O.B.) from 13d.p.c. embryos. After hybridisation, products were digested and run on a polyacrylamide gel.
Expression of the endogenous gene can be seen as a protected fragment of 156bp (A) in all the RNA samples (lanes 1-11). The GAPdH riboprobe is protected as 65 and 67bp fragments (B) and controls for the amount of total RNA in the hybridisation reaction.
M; size marker, ddT terminated sequencing of -40 primed bacteriophage M13mp18 control DNA.

Figure 6.6:


M
1 - R1 ES cells
2 - I114 ES cells
3 - 13d.p.c.(-/-) liver
4 - 13d.p.c.(-/-) head
5 - 13d.p.c.(-/-) R.O.B.
6 - 13d.p.c.(+/-) liver
7 - 13d.p.c.(+/-) head
8 - 13d.p.c.(+/-)R.O.B.
9 - 13d.p.c. wt liver
10 - 13d.p.c. wt head
11 - 13d.p.c. wt R.O.B.
156bp
67bp
65bp
A
B

(b) RT-PCR Analysis

Repeated screening of cDNA libraries with both the Group I and Group II fusion sequences has failed to identify the cDNAs containing the Group II fusion sequence. However, transcripts containing the Group II fusion sequence can be predicted from the relative position of the Group I and Group II fusion sequences in the genome and *gtar* sequence (Figure 6.7A). To assess if this predicted transcript exists and to analyse its expression pattern in the embryo, RT-PCR was performed on embryonic RNA samples using primers complementary to the predicted transcript (Figure 6.7A). RT-PCR was carried out on RNA from R1 ES cells, and the liver, head and R.O.B. from wild type 10.5d.p.c. embryos using primer LST-1 (complementary to the Group II fusion sequence) in combination with two different primers RTANK-2 and RTANK-3 (complementary to *gtar*; Figure 6.7A).

Both primer pairs amplified the predicted size product. Expression of this transcript in wild type tissues is restricted to the liver of mid-gestation in a similar pattern to the expression of the reporter gene in I114 embryos (Figure 6.7B lanes 1, 2, 5, and 6). No product was amplified in the absence of reverse transcriptase confirming that the products observed were not due to genomic contamination (Figure 6.7B lanes 10 and 12).

HPRT was used to control for the even loading of RNA in the RT-PCR reactions (Johansson and Wiles, 1995). The RT-PCR reactions were stopped after 25 and 30 PCR cycles to ensure that, in individual RNA samples, the RT-PCR amplification had not exhausted the reaction components. This is particularly important when using the HPRT primers. Individual samples, judged to be evenly loaded relative to other samples, could have reached an amplification plateau after fewer PCR cycles which would give the appearance of even RNA loading. The RT-PCR products were separated on an agarose gel and analysed by Southern blot hybridisation. End labelled primer LST-2 complementary to the Group II fusion sequence was used to probe for

Figure 6.7: Expression analysis of the endogenous Group II sequence using RT-PCR

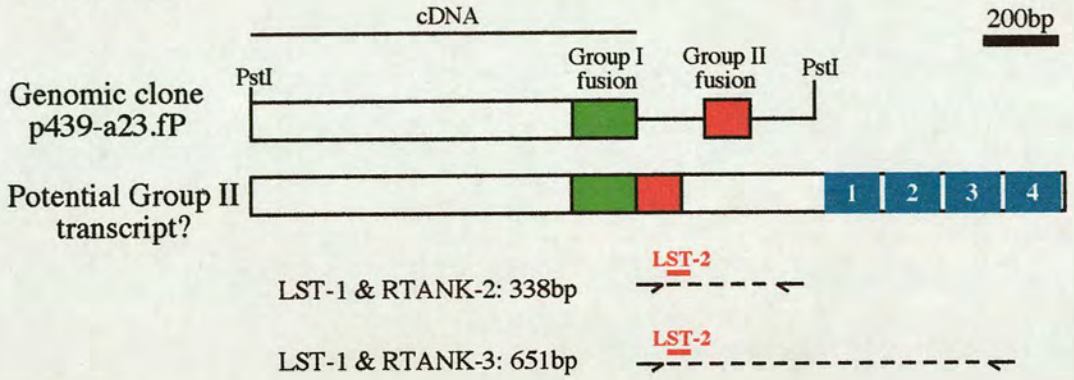
A: RT-PCR strategy

The sequence of the PstI fragment from 439-a23 places the *gtar* cDNA sequence from position 1-1017 which incudes the Group I sequence (green box) and the Group II fusion sequence (red box) within 200bp of each other.
From the genomic structure, a cDNA molecule containing the Group II fusion sequence can be predicted with the Group II sequence inserted between nucleotides 1017 and 1018. Primers were designed complementary to this hypothetical molecule. LST-1 to the Group II sequence and RTANK-2 and RTANK-3 to the endogenous cDNA sequence. The predicted PCR products of 338bp and 651 bp are shown which could be probed with oligo' LST-2 complementary to the Group II sequence.

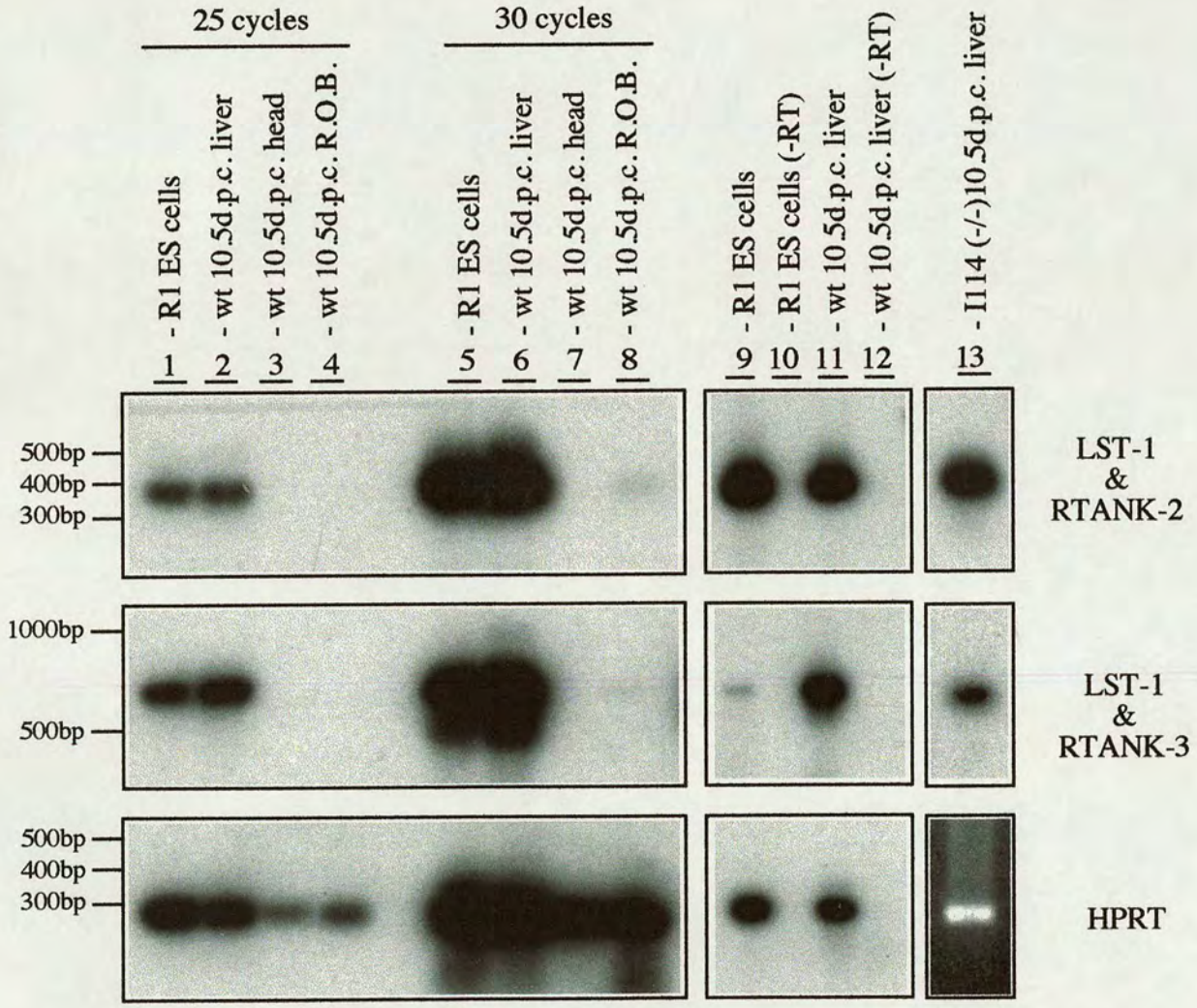
B: Results of RT-PCR using LST-1 with either RTANK 2 & RTANK-3.
The RT-PCR amplification products were separated on an agarose gel, Southern blotted and probed with end labelled LST-2 oligo'. The PCR products from each reaction were analysed after 25 PCR cycles (lanes 1-4) or 30 PCR cycles (lanes 5-8). The expected 381bp amd 651bp PCR products from the LST-1/RTANK-2 and the LST-1/RTANK-3 reactions respectively are only amplified a from R1 ES cells and wild type 10.5d.p.c. foetal liver after 25 and 30 cycles (lanes 1, 2, 5 and 6). No product is amplified from wild type 10.5d.p.c. head and R.O.B. (lanes 3, 4, 7 and 8).
Amplification of HPRT is used to control for the amount of total RNA used in the RT reaction (lanes 1-8).
To control for genomic DNA contamination, RT-PCR reactions were performed on RNA from R1 ES cells and wild type 10.5d.p.c. foetal liver with (lanes 9 and 11) without (lanes 10 and 12) reverse transcriptase.
Lane 13 shows that both primer sets amplify product from I114(-/-) 10.5d.p.c. RNA.

Figure 6.7:

A: RT-PCR strategy



B: RT-PCR



the predicted Group II containing transcript (Figure 6.7B). Primer 22 was used to probe for the HPRT amplification products.

The same three primer sets were used on RNA from 10.5d.p.c. I114 homozygous livers to assess if the integration of the gene trap vector had disrupted the production of the Group II containing transcript. Figure 6.7B (lane 13) shows that the Group II transcript is produced in I114 homozygous livers. The failure of the gene trap vector to disrupt the production of the Group II containing transcript in I114 homozygous tissues is consistent with what is observed for the endogenous Group I fusion transcript sequence (Figure 4.4) and the endogenous gene sequence downstream of the putative gene trap vector insertion site (Figure 6.6) in I114 homozygous tissues.

6.7. Cloning and Sequencing of the Group II Containing Transcript

The RT-PCR amplification products using the LST-1/RTANK-2 and LST-1/RTANK-3 primer sets were subcloned into pCRII-TOPO using the TOPO TA Cloning® kit (Invitrogen) and sequenced. The sequence of the RT-PCR products using primers LST-1 and RTANK-3 is shown in Figure 6.8. The splicing of the Group II sequence to the downstream *gtar* sequence is mediated by the same splice donor site used by the gene trap vector and predicted by the NIX database analysis of the Group II genomic sequence (position 49). Moreover, the splicing of the Group II sequence occurs via the same downstream endogenous splice acceptor used by the Group I sequence in *gtar* (position 50, Figure 6.8; position 1018, Figure 6.1). Conceptual translation of the Group II containing transcript identifies an ORF corresponding to the same reading frame predicted for the *gtar*. Translation of the endogenous Group II transcript in this frame corresponds to translation of the Group II fusion transcript in frame 2 (Figure 4.2). The sequence and conceptual translation of the endogenous Group II transcript will have predictable consequences on the protein produced from this transcript. The consequences will however be different depending

Figure 6.8: Sequence of the endogenous Group II transcript

The RT-PCR products from the LST-1/RTANK-2 and LST-1/RTANK-3 primer sets were subcloned and sequenced isolating the endogenous Group II sequence independent of the gene trap vector.

The primers used to amplify the 651bp Group II transcript are overlined.

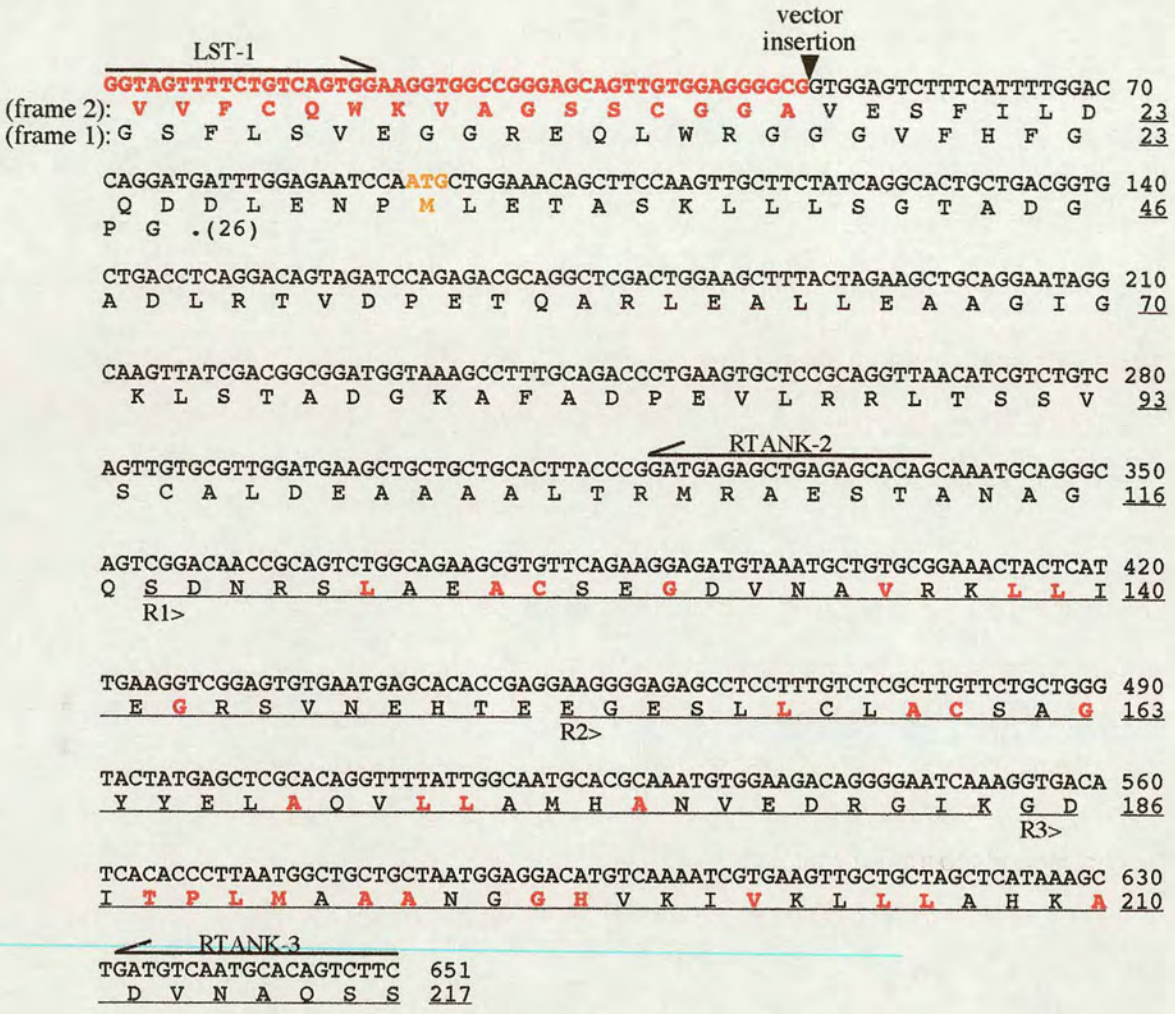
The Group II sequence (highlighted in red) splices to the endogenous downstream sequence via the splice donor used by the gene trap vector (49) and the splice acceptor used by the Group I sequence in the endogenous gene (50). This would put the vector insertion site between nucleotide 49 and 50 (black arrow).

Conceptual translation of the nucleotide sequence identifies an ORF which maintains the sequence of the endogenous cDNA and includes the first three ANK repeats (underlined and numbered). Frame 1 and 2 refers to the reading frame of the Group II sequence relative to the gene trap vector integration (Figure 4.2).

Frame 1 would be the reading frame if this transcript (via the Group II sequence) splices to the upstream Group I sequence as predicted in Figure 5.6B(i). Translation in this reading frame would terminate at amino acid position 26.

Frame 2 would potentially be the reading frame if this transcript is produced from a separate promoter immediately upstream of the Group II sequence. Consequently, the potential translational start codon is at amino acid residue 31 (highlighted in orange) which corresponds to amino acid residue 180 of the full length protein.

Figure 6.8:


vector
insertion
LST-1
GGTAGTTTTCTGTCAGTGGAAGGTGGCCGGGAGCAGTTGTGGAGGGGCGGTGGAGTCTTTCATTTTGGAC 70
(frame 2): V V F C Q W K V A G S S C G G A V E S F I L D 23
(frame 1): G S F L S V E G G R E Q L W R G G G V F H F G 23
CAGGATGATTTGGAGAATCCAATGCTGGAAACAGCTTCCAAGTTGCTTCTATCAGGCACTGCTGACGGTG 140
Q D D L E N P M L E T A S K L L L S G T A D G 46
P G .(26)
CTGACCTCAGGACAGTAGATCCAGAGACGCAGGCTCGACTGGAAGCTTTACTAGAAGCTGCAGGAATAGG 210
A D L R T V D P E T Q A R L E A L L E A A G I G 70
CAAGTTATCGACGGCGGATGGTAAAGCCTTTGCAGACCCTGAAGTGCTCCGCAGGTTAACATCGTCTGTC 280
K L S T A D G K A F A D P E V L R R L T S S V 93
RTANK-2
AGTTGTGCGTTGGATGAAGCTGCTGCTGCACTTACCCGGATGAGAGCTGAGAGCACAGCAAATGCAGGGC 350
S C A L D E A A A A A L T R M R A E S T A N A G 116
AGTCGGACAACCGCAGTCTGGCAGAAGCGTGTTCAGAAGGAGATGTAAATGCTGTGCGGAAACTACTCAT 420
Q S D N R S L A E A C S E G D V N A V R K L L I 140
R1>
TGAAGGTCGGAGTGTGAATGAGCACACCGAGGAAGGGGAGAGCCTCCTTTGTCTCGCTTGTTCTGCTGGG 490
E G R S V N E H T E E G E S L L C L A C S A G 163
R2>
TACTATGAGCTCGCACAGGTTTTATTGGCAATGCACGCAAATGTGGAAGACAGGGGAATCAAAGGTGACA 560
Y Y E L A Q V L L A M H A N V E D R G I K G D 186
R3>
TCACACCCTTAATGGCTGCTGCTAATGGAGGACATGTCAAAATCGTGAAGTTGCTGCTAGCTCATAAAGC 630
I T P L M A A A N G G H V K I V K L L L A H K A 210
RTANK-3
TGATGTCAATGCACAGTCTTC 651
D V N A Q S S 217

on whether the endogenous Group II transcript is expressed as an internal exon or from a separate promoter.

In Chapter 5, it was proposed that the Group II sequence could potentially splice to the upstream Group I fusion sequence via an upstream consensus splice acceptor sequence (Figure 5.4B). It was predicted that this splicing event would maintain the open reading frame of the Group I sequence (and by inference the ORF of the endogenous gene) by translation of the Group II sequence in frame 1 which contained no stop codons (Figure 5.4Bi). Moreover, such a splicing event and the translation of the Group II sequence in frame 1 would, in the context of gene trap vector splicing, result in the in-frame translation of *lacZ* and subsequently β-gal activity from this transcript. If the Group II transcript is translated in frame 1 as predicted, then splicing of this sequence to the downstream splice acceptor site results in the translation of the downstream endogenous gene sequence in a reading frame containing a stop codon 9 residues downstream of the translated Group II sequence (Figure 6.8). Consequently, if a protein is produced from this transcript, it will contain the first 65 amino acids from the upstream endogenous sequence (including the Group I sequence), 27 amino acids from the Group II sequence and a further 9 amino acids from the downstream endogenous sequence before termination. This would result in a 101 amino acid protein containing the two novel tandem repeat sequences, the hyperacidic cluster but lacking the downstream ANK repeat sequences. The liver specific expression of the Group II sequence (via an alternate liver specific splicing mechanism) would consequently result in this truncated protein isoform being expressed exclusively in the liver.

Another possible mechanism of the liver specific expression of the Group II sequence was via an alternative liver specific promoter. From the predicted transcriptional start site two nucleotides upstream of the Group II sequence, no translational start codons were found, suggesting that the Group II sequence would be untranslated in such a transcript. Consequently, translational initiation of the endogenous Group II transcript should occur downstream of the Group II sequence.

The first methionine codon in-frame with the same ORF of the full length cDNA is at position 92 (Figure 6.8). The use of this methionine codon for translational initiation would, relative to the full length protein sequence, produce a protein lacking the N-terminal 179 amino acid residues which includes the two novel repeat regions and the hyperacidic cluster. However, it is unclear if this is a genuine start codon as it is not in the context of a strong Kozak sequence necessary for translational initiation (Figure 6.8). In the context of splicing to the gene trap vector, the production of β-gal activity from the Group II fusion transcript transcribed from the predicted promoter would require translation of *lacZ* from its own start codon.

6.8. Discussion

Function of *gtar*

Translation of the *gtar* mRNA has identified a number of different protein motifs which may help to elucidate the cellular function of this gene. The most obvious are the 26 ANK repeat units which are present in two domains of 15 and 10 repeats. In addition, two potential PEST sequences and three potential bipartite NLS have also been identified.

Function of ANK Repeats

The ANK repeat was initially identified in the cell cycle regulatory proteins *cdc10* and *SW16* from *Scizosaccharomyces pombe* and *Saccharomyces cerevisiae* respectively. Subsequently, the ANK repeat motif has been identified in many different species, from *E.coli* to humans, in functionally diverse proteins including transcription factors (e.g. NFκ-B and GABPβ), cytoskeletal proteins (*Ank* 1, 2 and 3), and transmembrane proteins (e.g. *Notch*, *lin-12*) (Bork, 1993). The number of tandem ANK repeats varies between the different ANK repeat containing proteins. The majority of ANK repeat containing genes contain 3 to 8 ANK repeats. Outside of this group, *inversin* contains 16 ANK repeats, with the three mammalian ankyrins (*Ank1-3*) and α-latrotoxin from black widow spider venom containing 22 - 24 repeats (Bork, 1993; Morgan *et al.*, 1998). These repeat numbers compare to the 25 repeats identified in *gtar*. Despite the varied number of ANK repeats and the different functions of the ANK repeat containing genes, the ANK repeat itself is predicted to have a common role involved in mediating protein-protein interactions (Michaely and Bennett, 1992; Bork, 1993).

ANK Repeat Structure

The crystal structure of a number of ANK repeat containing proteins including $p18^{INK4c}$, 53BP2, and GABPβ has identified a common ANK repeat structure comprising essentially of a β-strand helix-turn-helix extended β-strand element (Wolberger, 1998; Figure 6.9D). The correlation between the ANK repeat residues and the different secondary structure elements is shown in Figure 6.9A.

The arrangement of the secondary structure units of the ANK repeat identifies essentially an L-shaped structure with the pairs of α-helices packing antiparallel to each other which, with adjacent ANK repeats, form α-helical bundles (stem of the L, Figure 6.9B). The first β-sheet of the ANK repeat packs antiparallell with the last β-sheet of the previous ANK repeat forming a β-hairpin structure in a perpendicular plane to the α-helices (base of the L, Figure 6.9B, Wolberger, 1998; Gorina and Pavletich, 1996; Venkataramani *et al.*, 1998). The highly conserved residues within the ANK repeat consensus play an important role in determining the folding of the secondary structure units. For example, the conserved glycine residues at position 2, 13, and 25 promote a sharp turn between the antiparallel β-strands, the two α-helices and the second α-helix and the adjacent β-strand respectively as well as terminating α-helices. Small hydrophobic residues are conserved at positions 5, 10, 18, 21 and 22 and promote intra- and inter-repeat packaging of the α-helices (Gorina and Pavletich, 1996; Venkataramani *et al.*, 1998). As well as the primary sequence of the ANK repeat, the folding of the ANK repeat secondary structure elements is also dependent on the presence of other ANK repeats, as extensive interaction between adjacent secondary structure elements stabilise the ANK repeat structure (Wolberger, 1998; Gorina and Pavletich, 1996; Venkataramani *et al.*, 1998). Consequently, the tertiary structure of ANK repeat regions is remarkably similar between different proteins (Figure 6.9D). An additional level of structural complexity has been postulated in proteins containing larger numbers of ANK repeats. *Ankl* has been shown to comprise four independently folded ANK repeat subdomains comprising of 6 ANK repeats each (Michaely and

Figure 6.9: Protein structure of ankyrin repeats

A: Diagram showing the relationship between the 33 amino acid residue ankyrin consensus and secondary structure elements. α1 and α2 are to two α-helices with β1 and β2 referring to two β-strands (adapted from Venkataramani *et al.*, 1998).
*: The ankyrin consensus is derived from 19 different ANK repeat containing proteins (Michaely and Bennett, 1992).

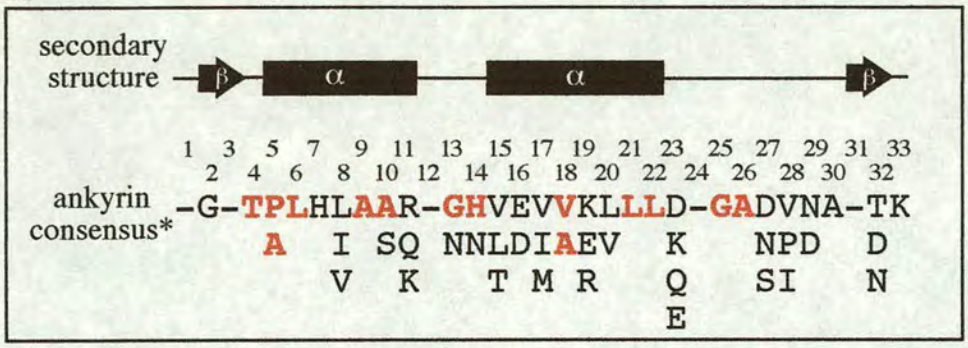
B: Topology diagram of the secondary structure elements of the 5 ANK repeat units from the cell cycle protein $p18^{INK4c}$.
The α-helices are depicted as circles (helical axis is perpendicular to the page) and the β-strands as arrows. The numbers refer to the amino acid residues of $p18^{INK4c}$ and are given at the beginning and end of each structural element. The third α-helix (α3) in the second ANK repeat is uncharacteristically short with a long loop between α3 and α4 (from Venkataramani *et al.*, 1998).

C: Overall structure of $p18^{INK4c}$ with the α-helical axis roughly perpendicular to the page and the β-strands parallel with the page (from Venkataramani *et al.*, 1998).

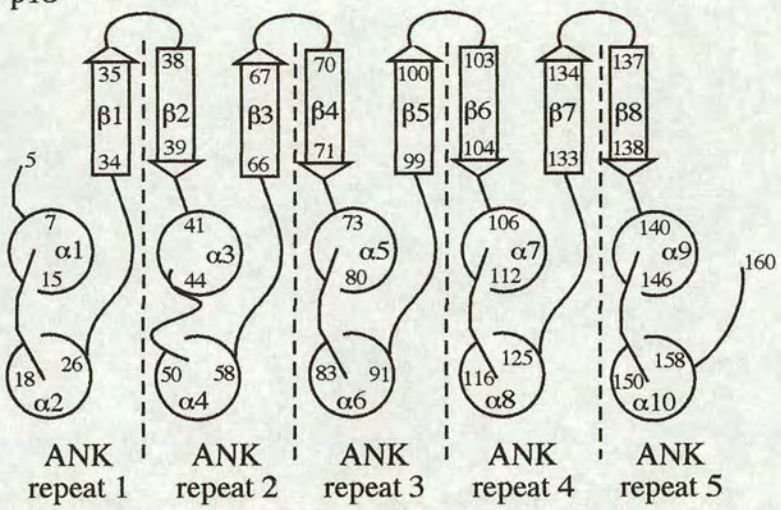
D: Comparison of the ANK repeat structure from $p18^{INK4c}$ and p53BP2.
ANK repeats 2-5 from $p18^{INK4c}$ are superimposed onto the 4 ANK repeats from p53BP. A remarkable similarity in tertiary structure is seen betweeen the two proteins (from Venkataramani *et al.*, 1998).

Figure 6.9:

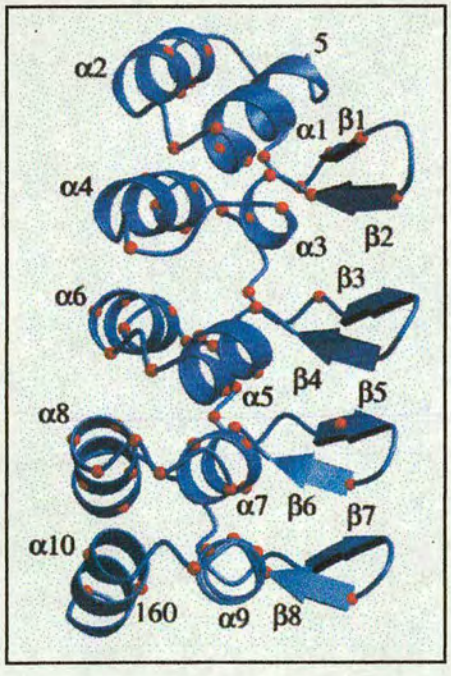
A



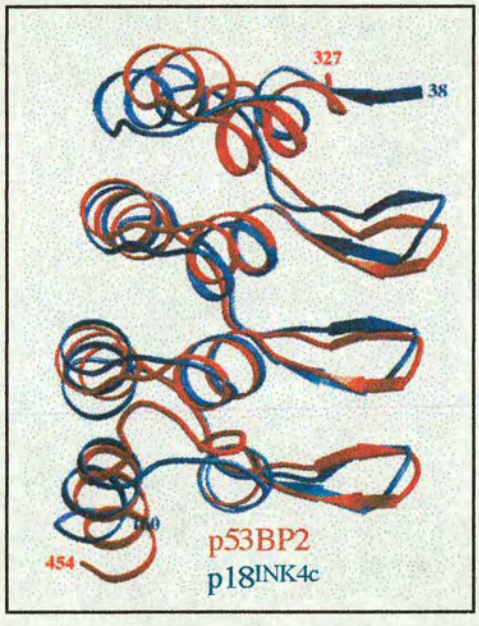
B: p18^{INK4c}



C: p18^{INK4c}



D



Bennett, 1993). It is, however, unknown if the six repeat sub-domains tertiary structure is the same as the other ANK repeat domains.

Although the highly conserved residues are important for ANK structure, they are not uniformly conserved in every ANK repeat only in selected ANK repeat units (Venkataramani *et al.*, 1998). This agrees with what is observed for *gtar* ANK repeat consensus where, for example, the glycine residue at position 2 is conserved in only 11 out of 26 repeats (Figure 6.2).

The ANK repeat motif binds to a range of target proteins from which no obvious sequence consensus or structural motif has been identfied (Bennett, 1992). The nature of the specificity of binding between the different ANK repeat containing genes and their different target proteins is potentially a consequence of variation within the ANK repeat consensus. The β-hairpin region, the solvent exposed α-helical regions and the large looped region connecting the second α-helix to the β-strand, generally lack sequence homology between ANK repeats. From crystal structure studies of both GABPβ and 53BP2, the β-hairpin region has been shown to be involved in interaction with their respective target molecules GABPα and p53 (Wolberger, 1998; Gorina and Pavletich, 1996).

The transcription factor GABP is a regulator of nuclear mitochondial genes functioning as a heterodimer consisting of the GABPα and GABPβ subunits. The GABPα subunit contains the ETS helix-turn-helix DNA binding domain which can bind DNA as a monomer. The GABPβ subunit contains 4 ANK repeats, functioning as the transactivation domain as well as increasing the affinity of GABPα DNA binding after dimerisation. All four ANK repeats within the GAPBβ subunit mediate dimerisation via the four β-hairpin structures (Wolberger, 1998).

Studies of other ANK repeat containing genes suggests that not every ANK repeat is equal in terms of their binding affinity to target molecules. The p53 binding protein 53BP2 contains 4 ANK repeats with 3 β-hairpin loops. Only the third β-hairpin structure is necessary for the interaction between 53BP2 and p53 (Gorina and Pavletich, 1996). Similarly, *Ank-1* (erythrocyte ankyrin) contains a 24 ANK repeat

domain which mediates high affinity binding to the membrane associated anion exchanger AE1. However, using defined regions of the ANK repeat domain, only repeats 21 and 22 were shown to be necessary for this high affinity binding. The ANK repeats not interacting with AE1 have the ability to bind other proteins including tubulin (Davis *et al.*, 1991).

From these studies it is apparent that the conserved residues within the ANK repeat consensus and the tandem arrangement of these repeats provides a scaffold for the interaction of the more variable residues within the ANK consensus (for example, the β-hairpin region) with the respective target proteins.

Functional Comparison Between *ankyrins* and *gtar*

The ANK repeat containing genes whose structure has been elucidated contain far fewer repeats than identified in *gtar* (26 ANK repeats). Ankyrins, which typically contain between 22-24 ANK repeats, have a more comparable repeat number. Moreover, the sequence of the ANK repeats from the endogenous trapped gene showed the highest homology to the ANK repeats found in ankyrin genes.

In mammals, 3 different ankyrin genes have been identified to date (*Ank1*, *Ank2* and *Ank3*) whose general function is to bind integral membrane proteins to the spectrin cytoskeleton (Lux *et al.*, 1990; Otto *et al.*, 1991; Peters *et al.*, 1995). Ankyrins are defined by a three domain structure comprising the N-terminal repeat domain, the spectrin binding domain and the C-terminal regulatory domain. The 89kD N-terminal domain contains 22-24 tandem ANK repeats and binds to a diverse array of membrane spanning molecules. The spectrin binding domain is approximately 62kD and is comprised of two sub-domains. The N-terminal sub-domain is highly acidic and is responsible for binding spectrin/fodrin. The C-terminal domain is overall basic in character with its sequence highly conserved between different ankyrins (Bennett, 1992; Peters *et al.*, 1995). The regulatory domain functions to modulate the binding affinity of the repeat domain and the spectrin binding domain. The sequence of the

regulatory domain is the least conserved region between the different ankyrins, a characteristic which is postulated to contribute to the different binding affinities of the ankyrin proteins (Bennett, 1992, Peters *et al.*, 1995). The ANK repeat domain of the ankyrin proteins has been shown to interact with a range of integral membrane proteins including the voltage-dependent Na^+ channel, the Na^+/Ca^{++} exchanger, CD44, IP3 receptor and Na^+/K^+-ATPase as well as the defining anion exchanger AE1 (Peters *et al.*, 1995). The role of ankyrin binding these ligands, with the exception of *Ank1* binding AE1 and spectrin in erythrocytes, is postulated to be in establishing and/or maintaining the polarized distribution of the integral membrane proteins in the cell (Peters *et al.*, 1995; Nelson and Veshnock, 1987; Hammerton *et al.*, 1991). Although *in vitro* the different ankyrins are capable of binding many of the same membrane proteins, the tissue restricted expression of the different ankyrin forms will presumably prevent them from doing so *in vivo* (Peters *et al.*, 1995).

It is highly unlikely that *gtar* corresponds to a novel ankyrin gene as no spectrin binding domain, which is highly conserved between ankyrins, was identified in the endogenous protein by motif searching or by homology to other proteins. Moreover, in ankyrins, the 24 ANK repeats are arranged in tandem. From studies on *Ank1*, these are folded into four separate subdomains of 6 ANK repeats. The full length *gtar* protein contains two separate ANK repeat domains of 15 and 10 repeats separated by 353 amino acid residues containing no obvious motifs or homologies. The splice variant SV-1, which has the majority of the sequence separating the ANK repeat domains deleted, still contains 59 amino acid residues between the separate ANK domains. The presence of either sequence separating the two ANK repeat domains is likely to preclude the formation of a similar sub-domain arrangement as observed for *Ank1*. Interestingly, the arrangement of the ANK repeats into two separate domains as observed in the endogenous gene has not been previously reported. Consequently, it would be of significant interest to determine which proteins acted as ligands for the I114 endogenous ANK repeat domain.

Common to both *gtar* and ankyrin genes is the use of alternative splicing to generate diversity from a single gene. All three ankyrin genes characterised to date display alternative splicing producing multiple protein isoforms (Lux *et al.*, 1990; Otto *et al.*, 1991; Peters *et al.*, 1995). For example, a minimum of seven different splice isoforms have been identified for the Ank3 gene (Peters *et al.*, 1995; De Matteis and Morrow, 1998). Several studies have identified functional differences between splice isoforms and their respective major ankyrin isoforms. For example, a splice variant of *Ank1* (2.2) lacks a 486 nucleotide sequence corresponding to a highly acidic amino acid sequence in the regulatory domain. The absence of this insert activates the ankyrin protein, enhancing both the binding of the ANK repeat domain to different membrane proteins and its spectrin binding affinity (Lux *et al.*, 1990). Interestingly, an alternative splice isoform of Ank3 has been identified which, although the amino acid sequence differs from Ank1, represents a deletion of a highly acidic region from the C-terminal regulatory domain. However, the functional significance of the *Ank3* splice isoform is not known (Peters *et al.*, 1995).

The identification of three potential splice isoforms of *gtar* may reflect a similar mechanism as observed for ankyrins where the splice isoforms alter the specificity of the ANK repeat domain binding to potential target proteins. For example, the splice isoform SV-2 contains only 13 ANK repeats and would be predicted to have a different binding specificity to the splice isoforms containing 25 repeats. This may be a consequence of the lower ANK repeat number and also the presence of the highly basic region containing the potential NLS immediately downstream of the 13 ANK repeats. Interestingly, a splice isoform of Ank1 has been identified containing a variable region of highly basic residues within the regulatory domain (Lambert *et al.*, 1990). However, it is unknown whether this alternative region has an effect on *Ank1* binding specificity.

The identification of a tissue specific isoform of the *Ank1* gene expressed from a separate promoter, is similar to the potential use of an liver specific alternative promoter by *gtar* to express the Group II sequence. The main *Ank1* protein isoform is

expressed as a 210kDa protein in erythrocytes and cerebellum (Bennett, 1992). A 25kDa isoform of *Ank1*, lacking the ANK repeats, has been identified which is expressed exclusively in skeletal muscle. Expression of this isoform is under the control of a muscle specific promoter located within intron 39 of the full length gene and expresses an alternative first exon which splices to three previously described *Ank1* exons (Birkenmeier *et al.*, 1998).

Identification of a Liver Specific Promoter?

Two potential mechanisms have been proposed as being responsible for the liver specific expression of the Group II sequence. Expression was either as a liver specific alternatively spliced internal exon or as the first exon of a transcript expressed from a liver specific promoter. Most of the evidence accumulated to date would suggest that the Group II sequence is expressed from a separate liver specific promoter.

RACE-PCR cloning from I114 RNA failed to isolate a fusion transcript containing the predicted splicing event between the Group I and Group II sequences (Figure 5.4B(i) & 5.5B(i)). The largest Group II RACE clone isolated from I114 RNA contained 58bp of endogenous sequence. From the predicted splicing event, the Group II sequence would extend a further 22 bases upstream to the putative splice acceptor sequence (Figure 5.4B). This sequence has never been amplified. The Group I and Group II sequences have never been isolated as a single transcript. These data are supported by the screening of an I114 12.5d.p.c. cDNA library with the *en-2* exon probe. Of 13 Group I positive cDNA clones isolated, none contained Group II sequence (data not shown). The 58bp of Group II sequence isolated from RACE is only 3 bases short of the transcriptional start site predicted from the putative promoter sequence (Figure 5.4C). Repeated attempts to clone larger Group II fusion transcripts by RACE-PCR using size selection and the use of a nested primer complementary to the Group II sequence proved unsuccessful. The failure to isolate sequence upstream of the Group II transcript suggests that this sequence may represent the 5' end of the

transcript as predicted from the promoter. Alternatively, this result may reflect a technical limitation of the RACE cloning protocol. Secondary structure within the postulated Group I and II transcript could be preventing reverse transcriptase from transcribing this sequence.

By examination of the sequence information, it is difficult to predict which model results in the expression of the Group II sequence. The predicted Group II splice acceptor sequence fits the consensus for a splice acceptor when the residues proximal to the splice site are examined and maintains the open reading frame initiated upstream. However, the Group II splice acceptor does not have a polypyrimidine stretch upstream of the proximal consensus sequence. Only 2 out of the 11 nucleotides immediately upstream of the proximal consensus are pyrimidines (Figure 5.4B), with only 7 pyrimidines present in the 40 nucleotides upstream (Appendix I). Whether this sequence functions as a splice acceptor is therefore unclear. Similarly, the promoter sequence identified by the TSSW algorithm does not contain an exact TATA box consensus sequence. Moreover, the three other promoter prediction algorithms in the NIX program failed to identify this sequence as being a promoter.

Conceptual translation of the endogenous Group II transcript identifies the same ORF as the full length cDNA. However, translation in this reading frame is incompatible with the reading frame of the Group I sequence in the context of the Group II sequence being an internal exon. This is, however, no absolute reason to eliminate the possibility of the Group II sequence existing as an internal exon. The role of the Group II sequence as the predicted internal exon could be to truncate the endogenous protein, eliminating the function of the ANK repeats.

The RT-PCR experiment described in Section 6.6 identified the endogenous Group II transcript using primers complementary to the Group II sequence and the downstream full length cDNA sequence. To assess if the Group II sequence is expressed as an internal exon, RT-PCR was performed using primers complementary to the Group II sequence and sequence potentially upstream in the endogenous full length cDNA. The predicted product would span the Group I sequence. In addition,

RT-PCR was performed to amplify the characterised endogenous cDNA using a primer upstream of the Group I sequence and the RTANK-2 primer complementary to the endogenous cDNA downstream of the Group I sequence. However, problems were encountered with both reactions in amplifying the expected product (data not shown). In the case of the Group II amplification reaction, this could be taken as an indication that the transcript containing the Group II sequence and the upstream sequences does not exist. However, the failure to amplify the expected product from the characterised endogenous cDNA suggests it is more likely to be a design limitation of the RT-PCR reaction. The results of the RACE cloning experiments suggests that amplification of the endogenous sequence upstream of the Group I sequence may be problematic with the largest Group I RACE clone isolated containing only 79bp of endogenous sequence. The failure to amplify upstream sequences may be a consequence of RNA secondary structure preventing the transcription of these sequences by reverse transcriptase .

Specific experiments can be designed to answer definitively whether expression of the Group II sequence is via alternate splicing or a separate promoter. S1 nuclease mapping would involve the design of a DNA probe spanning the predicted transcriptional start site. For example, the SfiI/PstI genomic DNA fragment from 439-a23.fP containing the Group II sequence would be ^{32}P-end labelled and hybridised to RNA from ES cells or foetal liver. The hybridisation products would then be digested with S1 nuclease which will digest only single stranded nucleic acid. The digestion products would be denatured, separated by polyacryamide gel electrophoresis to determine the size of the protected end labelled DNA molecule. From the size of the DNA, the 5' end of the Group II transcript can be defined.

Primer extension could also be used. A primer complementary to transcribed Group II sequence downstream of the predicted transcriptional start site would be designed, end labelled and hybridised to RNA. Extending from this primer, reverse transcriptase will duplicate the RNA molecule up to the 5' end of the transcript. The

products of this reaction would then be denatured and the radiolabelled DNA molecule sized on a polyacryamide gel to determine the start of the Group II transcript.

The activity of the predicted promoter could be assessed by fusing it upstream of a reporter gene (e.g β-geo). Conveniently, the 1.5Kb PstI genomic DNA fragment from PAC clone 439-a23 may contain sufficient promoter elements to drive reporter gene expression. The reporter expression pattern from such a construct could be monitored *in vitro* in ES cells and *in vivo* in chimaeric or transgenic animals to compare its expression with that of the gene trap integration event.

A Liver Specific Protein Isoform?

One of the most interesting aspects of the *gtar* gene trapped in the I114 line is the potential function of the liver specific Group II sequence. The liver specific expression of this sequence as either an alternatively spliced internal exon or as an alternative 5' end of the endogenous gene is predicted to produce different protein isoforms (Figure 6.10).

Expression of the Group II sequence as an internal exon is predicted to disrupt the ORF of the endogenous gene. This results in the production of a protein lacking the ANK repeats of the full length endogenous gene but retaining the novel repeat sequences and the hyperacidic cluster (Figure 6.10Bi). Consequently, the liver specific expression of the Group II transcript may serve to remove the protein binding ability of the endogenous gene but retain the function of the N-terminal region.

The expression of the Group II sequence as an alternative first exon of the endogenous gene driven by a liver specific promoter is predicted to produce a protein with a different N-terminus which lacks the two novel tandem repeat sequences and the hyperacidic cluster (Figure 6.10Bii). The functional consequences of this will depend on the role of the absent motifs. The two tandem repeats and the hyperacidic cluster are potentially incorporated within two PEST sequences. PEST sequences consist of regions rich in proline, glutamic acid, serine, threonine and, to a lesser extent, aspartic

Figure 6.10: Summary of Group I and Group II sequence expression in *gtar*

A: Splicing to the Group I fusion sequence

The Group I sequence is ubiquitously expressed from the endogenous promoter producing the full length gtar transcript and predicted protein.

B: Splicing of the Group II fusion sequence

(i) Group II sequence as an internal exon

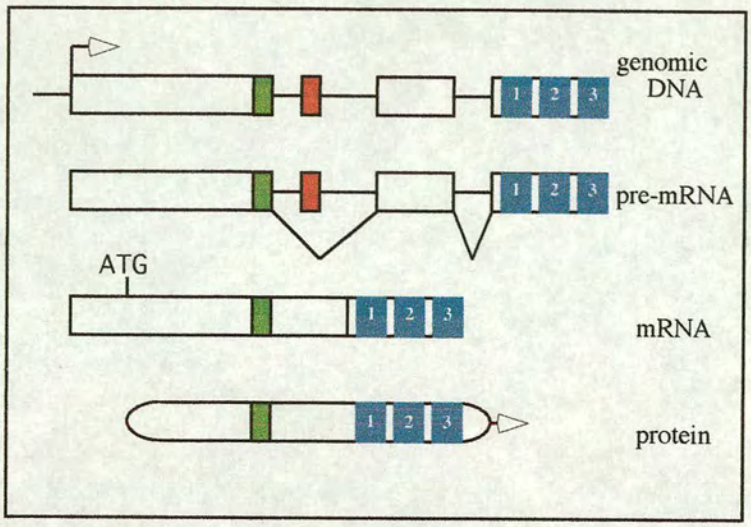
The Group I containing exon is ubiquitously expressed from the endogenous promoter. In the foetal liver, the Group II fusion sequence is alternatively spliced to downstream *gtar* sequences and the upstream Group I fusion sequence. The protein produced from this transcript will be initiated from the upstream start codon. This will place the *gtar* sequence downstream of the Group II sequence out-of-frame resulting in the production of a protein consisting of the N-terminus of predicted Gtar protein with no ANK repeats.

(ii) Group II sequence expressed from an alternative promoter

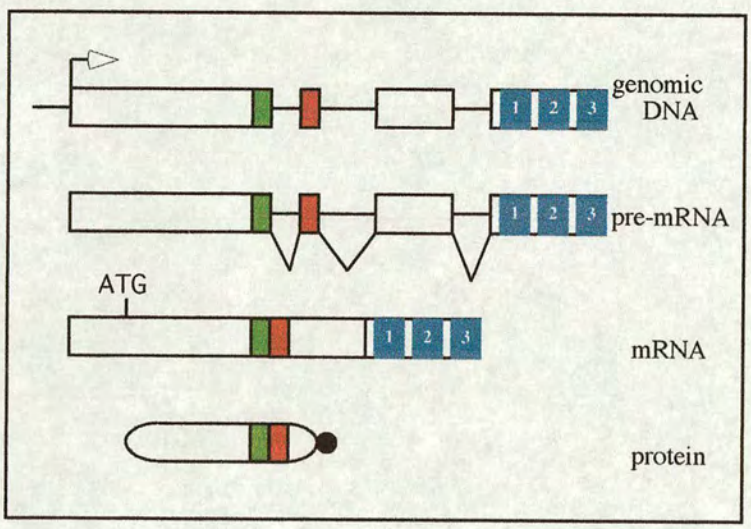
In the foetal liver, an alternative promoter drives expression of the Group II fusion sequence. In this transcript, the Group II sequence will be untranslated and a start codon within the *gtar* sequence used to initiate translation in the ORF predicted for the full length *gtar*. The protein produced will lack the N-terminus of the predicted full length Gtar protein.

Figure 6.10:

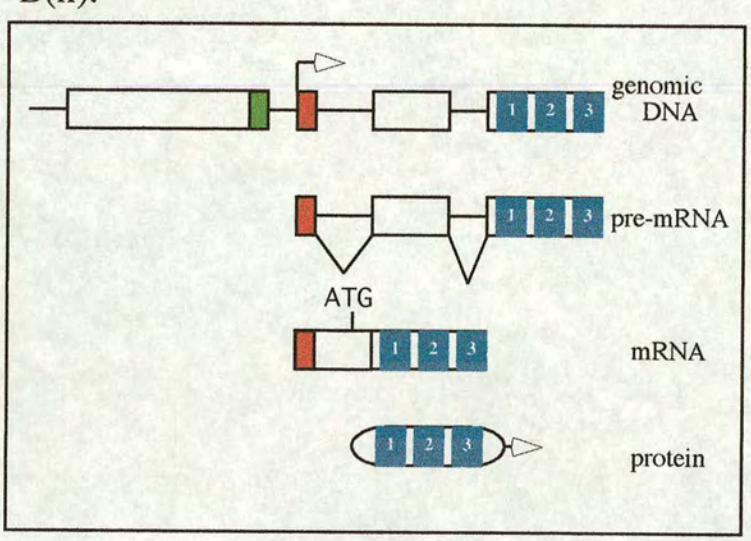
A:



B(i):



B(ii):



acid. Ten proteins shown to contain PEST sequences have intra-cellular half-lives of less than two hours while of 35 proteins with half-lives of between 20 and 220 hours, only two contained PEST sequences (Rogers *et al.*, 1986). If the N-terminus of the endogenous protein is functioning as a PEST sequence, then liver specific expression of the Group II sequence will remove this sequence resulting in a more stable ANK repeat protein whose expression is restricted to the foetal liver. Alternatively, the hyperacidic cluster may function, in the context of the full length gene, to modulate the binding specificity of the ANK repeats. A splice isoform of the *Ank1* gene lacking an acidic portion of the regulatory domain shows enhanced binding of the ANK repeat domain to integral membrane proteins (Lux *et al.*, 1990). Similarly, the absence of the hyperacidic cluster from the endogenous gene may enhance ANK repeat binding to target proteins in a liver specific manner.

It would also be of interest to investigate if the different *gtar* splice isoforms are expressed in a tissue specific manner similar to other ankyrin splice isoforms (Bennett, 1992; De Matteis and Morrow, 1998). Northern analysis, using probes specific to individual splice isoforms, or RT-PCR, using primers to amplify sequences which span the alternative splice sites used by the splice isoforms, will help to define their expression profile. The identification of a splice variant expressed exclusively in the foetal liver which is associated with the Group II sequence will help to further define a potential function of the liver specific *gtar* isoform.

The predicted outcomes of Group II transcript expression would significantly alter the properties of the endogenous protein produced (Figure 6.10B). The potential function of the liver specific GroupII transcript therefore could be to modulate the structure of the endogenous protein, either through the specificity of ANK repeat binding or protein stability, to perform a specific function necessary for hepatic specification and differentiation.

Chapter 7

CONCLUDING REMARKS

The aim of this project was to characterise the I114 gene trap cell line which shows restricted reporter activity in the foetal liver at mid-gestation. A detailed examination of the I114 reporter activity identified one of the earliest markers of hepatic specification which, unlike all previously analysed liver specific markers, is restricted to the definitive endoderm lineage. The desire to identify the endogenous gene responsible for such an exquisite expression pattern dominated the decision to characterise a complex gene trap integration event which did not result in an obvious phenotype. From the molecular characterisation, the lack of a phenotype is potentially a consequence of splicing around the gene trap vector by the endogenous gene. The liver specific expression pattern is a result of the gene trap vector splicing to a sequence present in a ubiquitously expressed gene which is differentially expressed exclusively in the foetal liver, either as an alternatively spliced exon or as the first exon of a liver specific promoter . Characterisation of the I114 integration does highlight certain aspects of gene trapping which contribute to the debate regarding the overall efficiency of gene trap vectors as a means of mutating the mouse genome. Moreover, it has identified a unique mechanism of gene trap vector action not previously identified which will facilitate the characterisation of future gene trap integrations.

The most obvious shortcoming of gene trapping identified in the I114 integration was the failure of the gene trap vector to function as an effective mutagen. Consequently, the identification of a functional role for the trapped *gtar* gene is unresolved. Although there is no direct evidence that exon trap vectors are more efficient mutagens than gene trap vectors, gene trap vectors were used in all of the entrapment events which do not result in a null allele (Table 1.1). One can predict that eliminating the need for a splicing event for entrapment removes the possibility of splicing around the gene trap vector. In using exon trapping, therefore, the initial

efficiency of entrapment would be reduced (Friedrich and Soriano, 1991) but the likelihood that the cell lines produced are functionally disrupted will perhaps be increased.

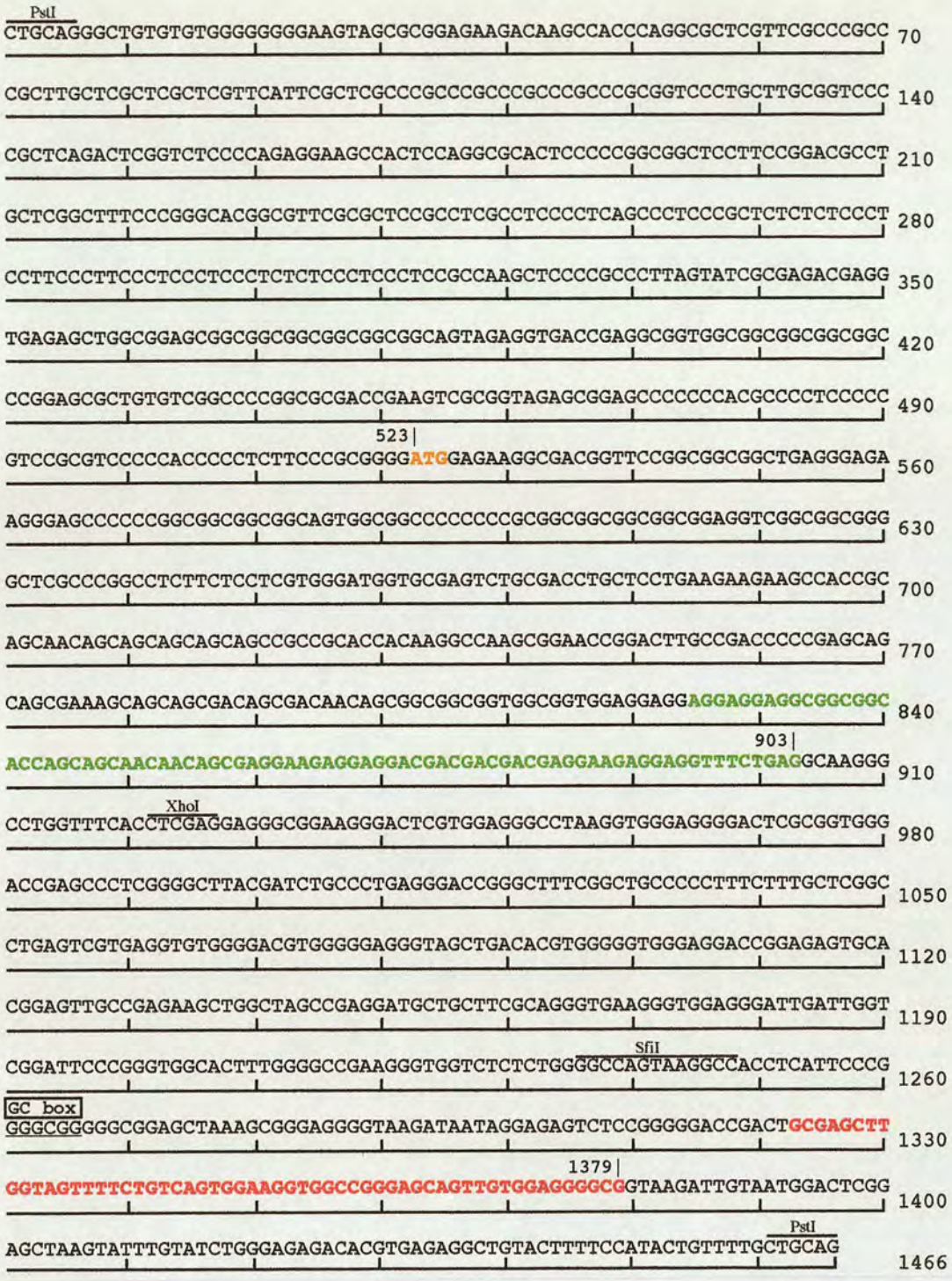
One of the most appealing aspects of gene trapping is the relative ease with which the trapped gene can be identified. In all the large scale gene trap screens carried out to date individual cell lines are prescreened to identify the trapped gene and complex integration events, such as multiple transcript production and intron containing lines (Hicks *et al.*, 1997; Chowdhury *et al.*, 1997; Townley *et al.*, 1997; Zambrowicz *et al.*, 1998). In more focused gene trap screens aimed at identifying developmental expression patterns or phenotypes, the interest of the investigator is determined by criteria other than the ease with which the trapped gene can be characterised. Such is the case with the I114 gene trap cell line. The extensive molecular characterisation of the I114 gene trap integration has identified alternate splicing to the gene trap vector not previously reported in the context of gene trapping. I114 therefore provides a model for the characterisation of multiple fusion transcript producing gene trap cell lines.

One of the potential benefits of gene trapping is the identification of cell type or tissue specific markers with which to study development. Even without the molecular data gleaned from the I114 line, the reporter activity provides an excellent marker for studying hepatic development which, due to the exclusivity of expression in the definitive endoderm, has benefits over existing hepatic markers.

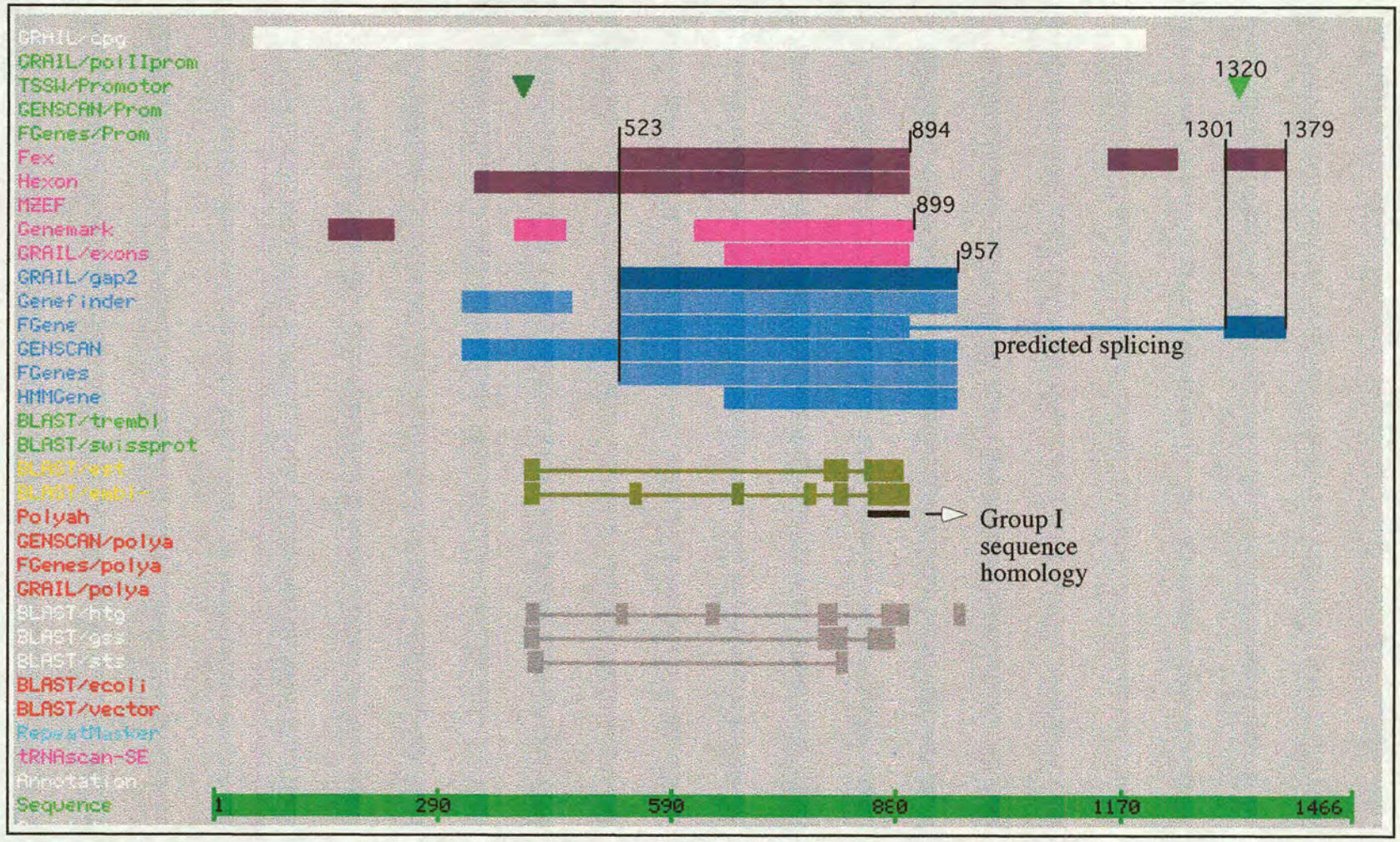
The I114 gene trap integration presents an excellent opportunity for further study. The function of *gtar* could be assessed using gene targeting technology. It would be necessary to disrupt the function of the Group II liver specific transcript separately from the ubiquitously expressed Group I sequence to identify functional differences between the two proteins. If the Group II sequence is expressed from a liver specific promoter, then the isolation of this sequence would provide a useful tool for sudying liver develoment. The promoter itself could be used to identify and isolate the transcription factors driving the expression of the Group II transcript. Similar studies on the promoter elements of liver specific genes, such as transthyretin, resulted

in the isolation of the essential hepatic transcription factors HNF-3 and HNF-4 (Costa *et al.*, 1989). Alternatively, the promoter could be used to drive the expression of specific molecules to the liver. For example, HNF-4 controls the expression of a number of liver specific factors. However, mice lacking HNF-4 die before the formation of the liver making it impossible to assess the role of HNF-4 in the liver (Chen *et al.*, 1994a). The Group II promoter could be used to drive expression of Cre recombinase in a targeted mouse line containing HNF-4 flanked by *loxP* sites which could potentially disrupt the function of the gene specifically in the foetal liver.

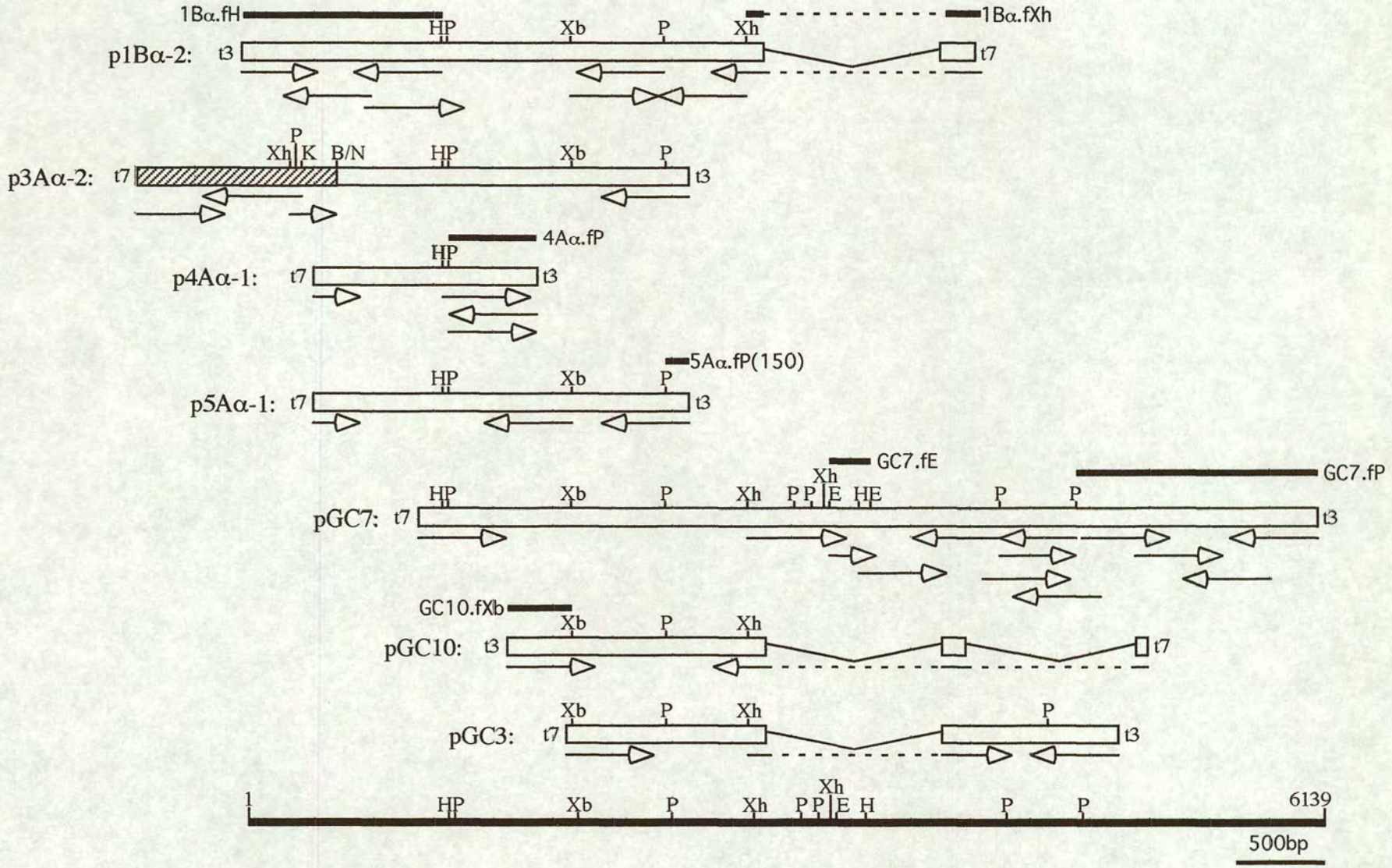
Despite the complex nature of the I114 gene trap integration, its characterisation has defined a novel mechanism of gene trap vector action and potentially identified a unique promoter activity restricted to the hepatic lineage of the definitive endoderm.



Appendix I: Sequence of 439-a23.fP
Sequence of the 1.5Kb genomic PstI fragement from PAC 439-a23. The sequence in red corresponds to the Group I fusion sequence and the sequence in green the Group II fusion sequence. Positions 903 & 1379 are the splice donor sites used by the gene trap vector. The ATG highlighted at position 523 corresponds to the start codon of the endogenous gene.



Appendix II: NIX analysis of 439-a23.fP
Diagrammatic output of the NIX suite of DNA analysis programs. The numbering corresponds to the nucleotide position on the positive strand of 439-a23.fP (same orientationas the fusion transcripts). Down the left hand side is listed the different DNA analysis programs run by NIX. The blocks represent a match between the entered sequence and the respective program.



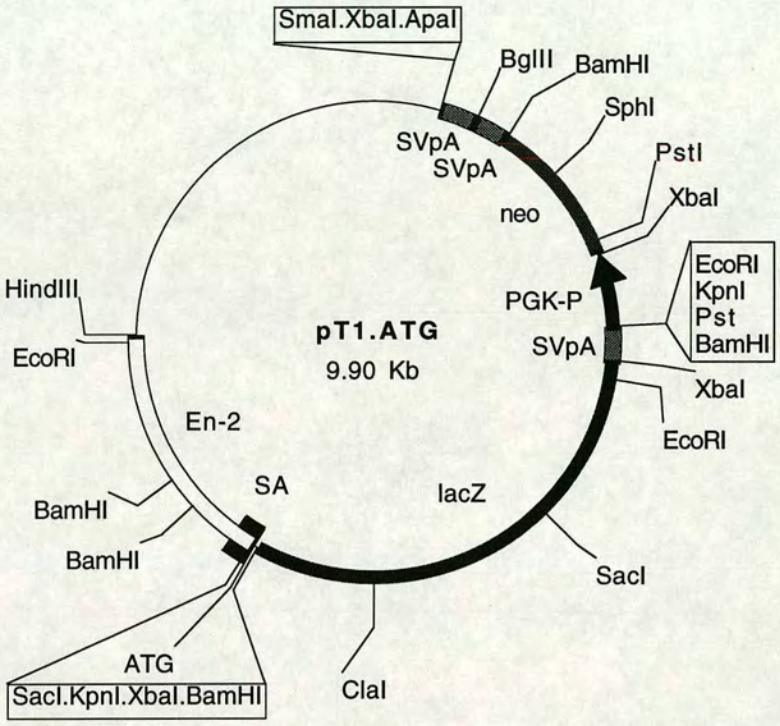
Appendix III: cDNA library clones

Restriction maps and sequencing strategy of the endogenous cDNA clones used to identify the full length cDNA and splice variants. All inserts are cloned into the EcoRI site of pBluescript SK(-). The orientation within pBluescript SK(-) is indicated by the t3 and t7 primer sites. The different probes used for RFLP analysis derived from these clones are identified by a solid black line above the restriction fragment. Hatched area corresponds to region of β-tubulin homology of hybrid cDNA clone p3Aα-2 H, HindIII; P, PstI; Xb, XbaI; Xh, XhoI, E, EcoRI.

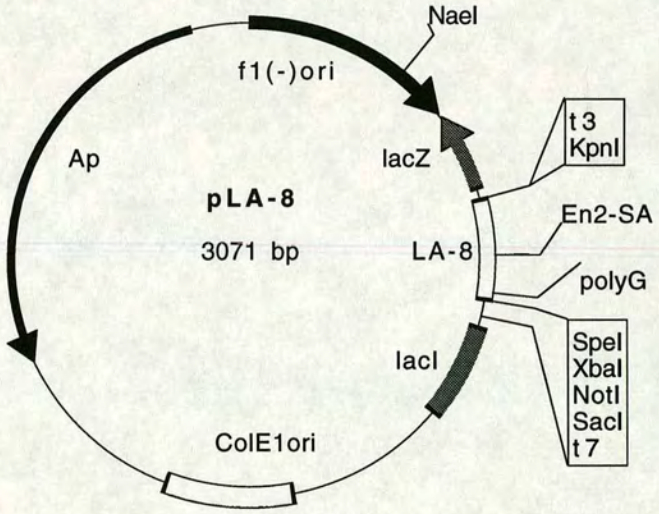
Appendix IV

Maps of the plasmid constructs used in the production of probes and riboprobes (created using MacVector V.5)

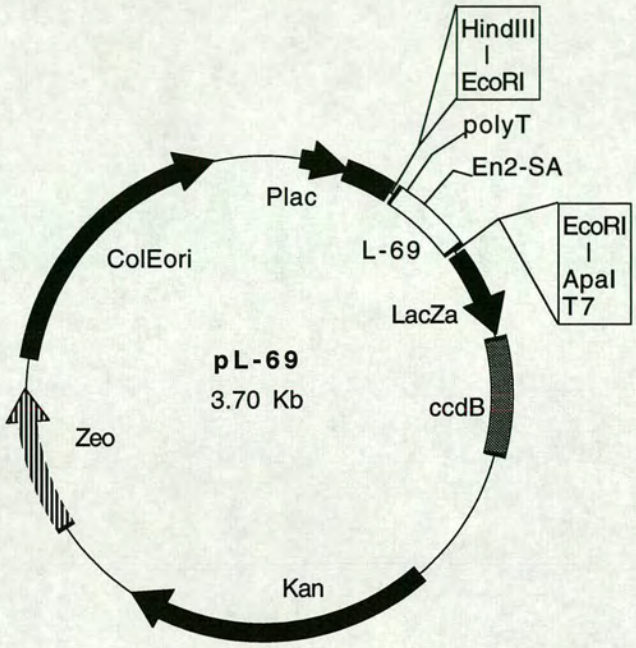
pT1-ATG



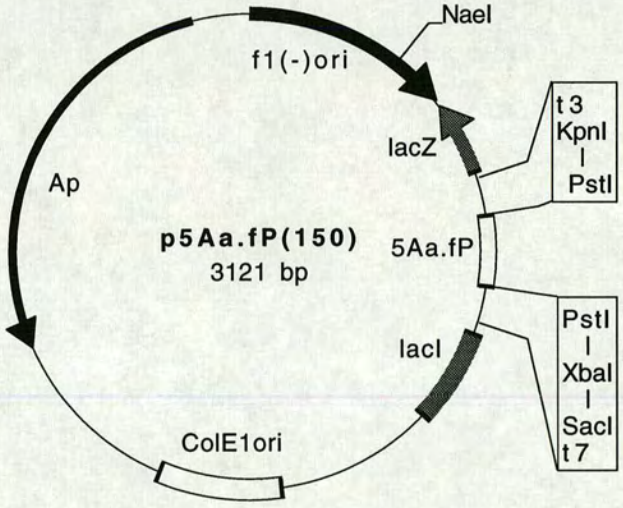
pLA-8



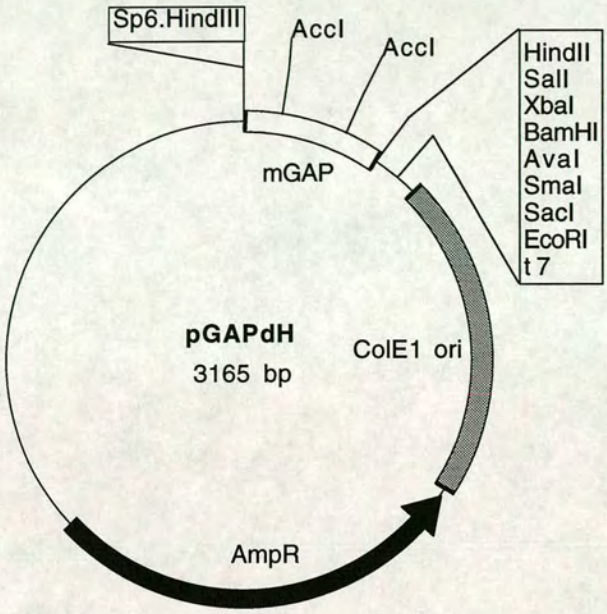
pL-69



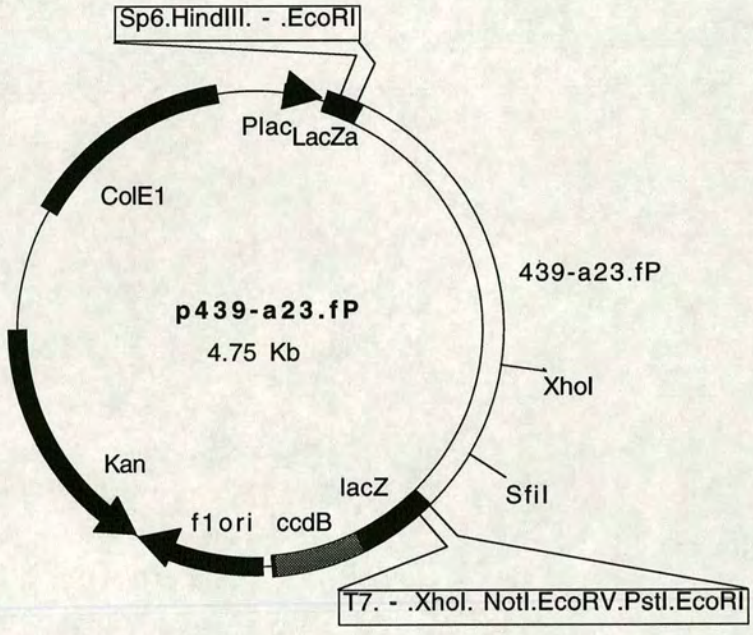
p5Aα.fP(150)



pGAPdH



p439-a23.fP



All the plasmids isolated from the cDNA library screening are cloned into the EcoRI site of pBluescript SK(-). The cDNA inserts and the probes derived from them are shown in Appendix III with the orientation indicated by the t3 and t7 primer sites.

Abbreviations

AFP	α-foetoprotein
alb	serum albumin
ANK	ankyrin-like repeat
AP	alkaline phosphatase
ATG	translational start codon
β-gal	β-galactosidase
β-geo	β-galactosidase and neomycin fusion
bp	basepairs
BSA	bovine serum albumin
cDNA	complementary deoxyribonucleic acid
c.p.m.	counts per minute
dATP	deoxyadenosine triphosphate
dCTP	deoxycytidine triphosphate
dGTP	deoxyguanosine triphosphate
dNTP	deoxynucleoside triphosphate
d.p.c.	days *post coitum*
dTTP	deoxythymidine triphosphate
DEN	diethylnitrosamine
DEPC	diethylpyrocarbonate
DIA	differentiation inhibiting activity
DMSO	dimethylsulfoxide
DNA	deoxyribonucleic acid
DTT	dithiothreitol
EB	embryoid body
EDTA	ethylenediaminetetraacetic acid
en-2	mouse engrailed-2
ENU	ethylitrosourea
ES cells	embryonic stem cells
EST	expressed sequence tag
ER	endoplasmic reticulum
FISH	fluorescent *in situ* hybridisation
FITC	flourescein isothiocyanate
GFP	green flourescent protein
gtar	gene trap ankyrin repeat gene
HNF	hepatocyte nuclear factor
hsp	heat shock protein

IRES	internal ribosomal entry site
kb	kilobases
LIF	leukemia inhibiting factor
LTR	long terminal repeat
mRNA	messenger ribonucleic acid
NLS	nuclear localisation signal
ORF	open reading frame
PAC	p1 artificial chromosome
PCR	polymerase chain reaction
PFGE	pulse field gel electrophoresis
PGK	phophoglycerate kinase
polyA	polyadenylation
RA	retinoic acid
RACE	rapid amplification of cDNA ends
RFLP	restriction fragment length polymorphism
RNA	ribonucleic acid
RPA	RNase protection assay
RT-PCR	reverse transcription polymerase chain reaction
SA	splice acceptor
SD	splice donor
SF/HGF	scatter factor/ hepatocyte growth factor
SV	splice variant
TESPA	3-aminopropyltriethoxysilane
UTR	untranslated region

List of Figures

Figure 1.1: Mode of Action of the Two Basic Entrapment Vector Systems
Figure 1.2: Promoter Trapping
Figure 1.3: PolyA Trapping
Figure 1.4: Gene Trapping Using Site Specific Recombination
Figure 1.5: Model of Secretory Trap Vector Action
Figure 1.6: Inductive Interactions During the Formation of the Foetal Liver

Figure 2.1: Overview of 5'RACE-PCR (Protocol 1)

Figure 3.1: I114 Reporter Activity During Early Liver Development
Figure 3.2: I114 Reporter Activity from Mid-Late Gestation
Figure 3.3: I114 Adult Reporter Activity
Figure 3.4: Comparison of I114 and AFP Reporter Activity During Embryogenesis
Figure 3.5: $AFP^{(\beta\text{-geo}/+)}$ Adult Reporter Activity
Figure 3.6: I114 Reporter Activity During Tumourigenesis
Figure 3.7: Southern Blot Analysis of I114 Genomic DNA

Figure 4.1: Analysis of the I114 Fusion Transcript by Northern Blot
Figure 4.2: I114 Fusion Transcript Sequences
Figure 4.3: Direct Sequence of I114 Fusion Transcripts
Figure 4.4: Expression Analysis of Group I Fusion Transcript LA-8 by RNase Protection
Figure 4.5: Expression Analysis of Group II Fusion Transcript L-69 by RNase Protection
Figure 4.6: Expression Analysis of Group II Fusion Transcript L-69 by RT-PCR
Figure 4.7: β-galactosidase Protein Expression in I114 Embryos
Figure 4.8: Mechanisms of Multiple Fusion Transcript Production

Figure 5.1: Chromosomal Mapping of the I114 Gene Trap Integration
Figure 5.2: PFGE Analysis of I114 Genomic DNA
Figure 5.3: Characterisation of Genomic DNA Containing Group I and Group II Sequences
Figure 5.4: Analysis of 439-a23.fP Genomic Sequence
Figure 5.5: Summary of the I114 Gene Trap Integration Event

Figure 6.1: Nucleotide and Protein Sequence of *gtar*
Figure 6.2: Optimal Alignment of the *gtar* ANK Repeats

Figure 6.3: Identification of a Novel Tandem Repeat Sequence
Figure 6.4: Splice Variants of *gtar*
Figure 6.5: Genomic Structure of *gtar*
Figure 6.6: Expression Analysis of *gtar* by RNase Protection
Figure 6.7: Expression Analysis of the Endogenous Group II Sequence by RT-PCR
Figure 6.8: Sequence of the Endogenous Group II Transcript
Figure 6.9: Protein Structure of Ankyrin Repeats
Figure 6.10: Summary of Group I and Group II Sequence Expression in *gtar*

List of Tables

Table 1.1: Characterised Entrapment Events
Table 1.2: Sequence Data from 3 Large Scale Entrapment Screens

Table 3.1: Comparison of I114(+/-) and AFP$^{(\beta\text{-geo}/+)}$ β-gal Activity
Table 3.2: Genotype of I114 Heterozygous Intercross Litters at Weaning

Table 4.1: Number and Size of the Different RACE Clones Isolated From the I114 Gene Trap Line

References

Abelev, G. I. (1971). Alpha-fetoprotein in oncogenesis and its association with malignant tumours. Advances in Cancer Research *14*, 295- 358.

Adams, M. D., Kerlavage, A. R., Fleischmann, R. D., Fuldner, R. A., Bult, C. J., Lee, N. H., Kirkness, E. F., Weinstock, K. G., Gocayne, J. D., and White, O. (1995). Initial assessment of human gene diversity and expression patterns based on 83 million nucleotides of cDNA sequence. Nature *377*. 3-174.

Adams, M. D., Rudner, D. R., and Rio, D. C. (1996). Biochemistry and regulation of pre-mRNA splicing. Current Opinion in Cell Biology *8*, 331-339.

Alam, J., and Cook, J. L. (1990). Reporter genes: application to the study of mammalian gene transcription. Annals of Biochemistry *188*, 245-254.

Alberts, B., Bray, D., Lewis, J., Raff, M., Roberts, K., and Watson, J. D. (1994). Molecular Biology of the Cell, 3 Edition (New York & London: Garland Publishing, Inc.).

Alison, M. (1998). Liver stem cells: a two compartment system. Current Opinion in Cell Biology *10*, 710-715.

Allen, J. D., Lints, T., Jenkins, N. A., Copeland, N. G., Strasser, A., Harvey, R. P., and Adams, J. A. (1991). Novel murine homeobox gene on chromosome 1 expressed in specific hematopoietic lineages during embryogenesis. Genes and Development *5*, 509-520.

Allen, N. D., Cran, D. G., Barton, S. C., Hettle, S., Reik, W., and Azim Surani, M. (1988). Transgenes as probes for active chromosomal domains in mouse development. Nature *333*, 852- 855.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. Journal of Molecular Biology *215*, 403-410.

Ang, S.-L., and Rossant, J. (1994). *HNF-3b* is essential for node and notochord formation in mouse development. Cell *78*, 561-574.

Ang, S.-L., Wierda, A., Wong, D., Stevens, K. A., Cascio, S., Rossant, J., and Zaret, K. S. (1993). The formation and maintenance of the definitive endoderm lineage in the mouse: involvement of HNF3/*forkhead* proteins. Development *119*, 1301-1315.

Angel, P., Imagawa, M., Chiu, R., Stein, B., Imbra, R. J., Rahmsdorf, H. J., Jonat, C., Herrlich, P., and Karin, M. (1987). Phorbol ester-inducible genes contain a common *cis* element recognized by a TPA-modulated *trans* acting factor. Cell *49*, 729-739.

Arceci, R. J., King, A. A. J., Simon, M. C., Orkin, S. H., and Wilson, D. B. (1993). GATA-4: A retinoic acid-inducible GATA-binding transcription factor expressed in endodermally derived tissues and heart. Molecular Cell Biology *13*, 2235-2246.

Auer, K. L., Contessa, J., Brenz-Verca, S., Pirola, L., Rusconi, S., Cooper, G., Abo, A., Wymann, M. P., Davis, R. J., Birrer, M., and Dent, P. (1998). The Ras/Rac1/Cdc42/SEK/JNK/c-Jun cascade is a key pathway by which agonists stimulate DNA synthesis in primary cultures of rat hepatocytes. Molecular Biology of the Cell *9*, 561-573.

Ausubel, F. A., Brent, R., Kingston, R. E., Moore, D. D., Seidman, J. G., and Struhl, K. (1998). Current Protocols in Molecular Biology (New York: John Wiley & Sons).

Avner, P., Amar, L., Dandolo, L., and Guenet, J. L. (1988). Genetic analysis of the mouse using interspecific crosses. Trends in Genetics *4*, 18- 23.

Baeuerle, P. A., and Henkel, T. (1994). Function and activation of NF-kB in the immune system. Annual Review of Immunology *12*, 141-179.

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

Balling, R., Deutsch, U., and Gruss, P. (1988). *undulated*, a mutation affecting the development of themouse skeleton, has a point mutation in the paired box of *Pax-1*. Cell *55*, 531-535.

Barker, D. D., Wu, H., Hartung, S., Breindl, M., and Jaenisch, R. (1991). Retrovirus-induced insertional mutagenesis: mechanism of collagen mutation in Mov13 mice. Molecular Cell Biology *10*, 5154-5163.

Bedell, M. A., Jenkins, N. A., and Copeland, N. G. (1997). Mouse models of human disease. Part I: Techniques and resources for genetic analysis in mice. Genes and Development *11*, 1- 10.

Beg, A. A., Sha, W. C., Bronson, R. T., Ghosh, S., and Baltimore, D. (1995). Embryonic lethality and liver degeneration in mice lacking the RelA component of NF-kB. Nature *376*, 167- 170.

Bellen, H. J., O'Kane, C. J., 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 and Development *3*, 1288- 1300.

Bellen, H. J., Wilson, C., and Gehring, W. J. (1990). Dissecting the complexity of the nervous system by enhancer detection. BioEssays *12*, 199- 204.

Bellofatto, V., Shapiro, L., and Hodgson, D. A. (1984). Generation of a Tn5 promoter probe and its use in the study of gene expression in *Caulobacter crescentus*. Proceedings of the National Academy of Science, USA *81*, 1035-1039.

Bennett, V. (1992). Ankyrins. Journal of Biological Chemistry *267*, 8703- 8706.

Betto, H., Kaneda, K., Yamamoto, T., Kojima, A., and Sakurai, M. (1996). Development of intralobular bile ductules after spontaneous hepatitis in Long-Evans mutant rats. Laboratory Investigation *75*, 43-53.

Bhat, K., McBurney, M. W., and Hamada, H. (1988). Functional cloning of mouse chromosomal loci specifically active in embryonal carcinoma stem cells. Molecular and Cellular Biology *8*, 3251- 3259.

Bier, E., Vaessin, H., Shepherd, 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 and Development *3*, 1273- 1287.

Birchmeier, C., and Birchmeier, W. (1993). Molecular aspects of mesenchymal-epithelial interactions. Annual Review of Cell Biology *9*, 511-540.

Birkenmeier, C. S., J., S. J., Gifford, E. J., Deveau, S. A., and Barker, J. E. (1998). An alternative first exon in the distal end of the erythroid ankyrin gene leads to production of a small isoform containing an NH2-terminal membrane anchor. Genomics *50*, 79- 88.

Bladt, F., Riethmacher, D., Isenmann, S., Aguzzi, A., and Birchmeier, C. (1995). Essential role for the c-*met* receptor in the migration of myogenic precursor cells into the limb bud. Nature *376*, 768-771.

Boguski, M. S., and Lowe, T. M. (1993). dbEST- database for "expressed sequence tags". Nature Genetics *4*, 332-333.

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

Bork, P. (1993). Hundreds of ankyrin-like repeats in functionally diverse proteins: mobile modules that cross phyla horizontally? PROTEINS: Structure, Function, and Genetics *17*, 363-374.

Bossard, P., and Zaret, K. S. (1998). GATA transcription factors as potentiators of gut endoderm differentiation. Development *125*, 4909- 4917.

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

Bralet, M.-P., Calise, D., Brechot, C., and Ferry, N. (1996). In vivo cell lineage analysis during chemical hepatocarcinogenesis using retroviral-mediated gene transfer. Laboratory Investigation *74*, 871-881.

Brandon, E. P., Idzerda, R. L., and McKnight, G. S. (1995). Targeting the mouse genome: a compendium of knock-outs (Parts I-III). Curr Biol *5*, 625-634, 758-765, 873-1073.

Breathnach, R., and Chambon, P. (1981). Organisation and expression of eukaryotic split genes coding for proteins. Annual Review of Biochemistry *50*, 349- 383.

Brennan, J. (1997). A modified gene trap approach to identify secretory molecules involved in mouse development (Edinburgh: University of Edinburgh), pp. 151.

Brenner, D. G., Lin-Chao, S., and Cohen, S. N. (1989). Analysis of mammalian cell genetic regulation *in situ* by using retrovirus-derived "portable exons" carrying the Escherichia coli *lacZ* gene. Proceedings of the National Academy of Science, USA *86*, 5517-5521.

Brenner, T., Beyth, Y., and Abramsky, O. (1980). Inhibitory effect of a-fetoprotein on the binding of myasthenia gravis antibody to acetylcholine receptor. Proceedings of the National Academy of Science, USA *77*, 3635-3639.

Bullock, S. L., Fletcher, J. M., Beddington, R. S. P., and Wilson, V. A. (1998). Renal agenesis in mice homozygous for a gene trap mutation in the gene encoding heparan sulfate 2-sulfotransferase. Genes and Development *12*, 1894-1906.

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

Capecchi, M. R. (1989). Altering the genome by homologous recombination. Science *244*, 1288-1292.

Casabadan, M. J., and Cohen, S. N. (1979). Lactose genes fused to exogenous promoters in one step using a Mu-lac bacteriophage: *In vivo* control probe for transcriptional control sequences. Proceedings of the National Academy of Science, USA *76*, 4530-4533.

Cascio, S., and Zaret, K. S. (1991). Hepatocyte differentiation initiates during endodermal- mesenchymal interactions prior to liver formation. Development *113*, 217-225.

Chen, H., Egan, J. O., and Chiu, J.-F. (1997). Regulation and activities of a-fetoprotein. Critical Reviews in Eukaryotic Gene Expression *7*, 11-41.

Chen, J., Nachabah, A., Scherer, C., Ganju, P., Reith, A., Bronson, R., and Ruley, H. E. (1996). Germ-line inactivation of the murine Eck receptor tyrosine kinase by gene trap retroviral insertion. Oncogene *12*, 979-988.

Chen, W. S., Manova, K., Weinstein, D. C., Duncan, S. A., Plump, A. S., Prezioso, V. R., Bachvarova, R. F., and Darnell Jr, J. E. (1994a). Disruption of the HNF-4 gene, expressed in visceral endoderm, leads to cell death in embryonic ectoderm and impaired gastrulation of mouse embryos. Genes and Development *8*, 2466- 2477.

Chen, Z., Friedrich, G. A., and Soriano, P. (1994b). Transcriptional enhancer factor -1 disruption by a retroviral gene trap leads to heart- defects and embryonic lethality in mice. Genes and Development *8*, 2293-2301.

Chomczynski, P., and Sacchi, N. (1987). Single-step method of RNA isolation by acid guanidiniumthiocyanate-phenol-chloroform extraction. Analytical Biochemistry *162*, 156-159.

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

Church, G. M., and Gilbert, W. (1984). Genomic sequencing. Proceedings of the National Academy of Science, USA *81*, 1991-1995.

Conlon, F. L., Barth, K. S., and Robertson, E. J. (1991). A novel retrovirally induced embryonic lethal mutation in the mouse: assessment of the developmental fate of embryonic stem cells homozygous for the 413.d proviral integration. Development *111*, 969- 981.

Conlon, R. A., and Rossant, J. (1992). Exogenous retinoic acid rapidly induces anterior ectopic expression of murine *Hox-2* genes *in vivo*. Development *116*, 357-368.

Costa, R. H., Grayson, D. R., and Darnell, J. E. (1989). Multiple hepatocyte-enriched nuclear factors function in the regulation of transthyretin and a1-antitrypsin genes. Molecular and Cellular Biology *9*, 1415-1425.

Couldrey, C., Carlton, M. B. L., Ferrier, J., Colledge, W. H., and Evans, M. J. (1998). Disruption of murine a-enolase by a retroviral gene trap results in early embryonic lethality. Developmental Dynamics *212*, 284-292.

Davis, L. H., Otto, E., and Bennet, V. (1991). Specific 33 residue repeat(s) of erythrocyte ankyrin associate with the anion exchanger. Journal of Biological Chemistry *266*, 11163-11169.

De Matteis, M. A., and Morrow, J. S. (1998). The role of ankyrin and spectrin in membrane transport and domain formation. Current Opinion in Cell Biology *10*, 542-549.

Degregori, T., Russ, A., Vonmelchner, H., Rayburn, H., Priyaranjan, P., Jenkins, N. A., Copeland, N. G., and Ruley, H. E. (1994). A murine homolog of the yeast rna1 gene is required for postimplantation development. Genes and Development *8*, 265-276.

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

Dietrich, W. F., Miller, J., Steen, R., Merchant, M. A., and et al. (1996). A comprehensive genetic map of the mouse genome. Nature *380*, 149- 152.

Dingwall, C., and Laskey, R. A. (1998). Nuclear import: A tale of two sites. Current Biology *8*, R922-R924.

Dobie, K., Mehtali, M., McClenaghan, M., and Lathe, R. (1997). Variegated gene expression in mice. Trends in Genetics *13*, 127-130.

Doolittle, D. P., Davisson, M. T., Guidi, J. N., and Green, M. C. (1996). Catalog of mutant genes and polymorphic loci, 3rd Edition, M. F. Lyon, S. Rastan and S. D. M. Brown, eds. (Oxford: Oxford University Press).

Driever, W., SolnicaKrezel, L., Schier, A. F., Neuhauss, S., Malicki, J., Stemple, D. L., Stainier, D., Zwartkruis, F., Abdelilah, S., Rangini, Z., Belak, J., and Boggs, C. (1996). A genetic screen for mutations affecting embryogenesis in zebrafish. Development *123*, 37-46.

Drinkwater, N. R., and Ginsler, J. J. (1986). Genetic control of hepatocarcinogenesis in C57BL/6J and C3H/HeJ inbred mice. Carcinogenesis *7*, 1701- 1707.

Duncan, S. A., Manova, K., Chen, W. S., Hoodless, P., Weinstein, D. C., Bachvarova, R. F., and Darnell Jr, J. E. (1994). Expression of transcription factor HNF-4 in the extraembryonic endoderm, gut, and nephrogenic tissue of the developing mouse embryo: HNF-4 is a marker for primary endoderm in the implanting blastocyst. Proceedings of the National Academy of Science, USA *91*, 7598- 7602.

Dunwoodie, S. L., Rodriguez, T. A., and Beddington, R. S. P. (1998). *Msg1* and *Mrg1*, founding members of a gene family, show distinct patterns of gene expression during mouse embryogenesis. Mechanisms of Development *72*, 27-40.

Dziadek, M., and Adamson, E. (1978). Localisation and synthesis of alphafoetoprotein in post-implantation mouse embryos. Journal of Embryology and Experimental Morphology *43*, 289-313.

Dziadek, M. A., and Andrews, G. K. (1983). Tissue specificity of alpha-fetoprotein messenger RNA expression during mouse embryogenesis. EMBO Journal *2*, 549-554.

Dzierzak, E., and Medvinsky, A. (1995). Mouse embryonic hematopoiesis. Trends in Genetics *11*, 359- 365.

Evans, M. (1998). Gene trapping-A preface. Developmental Dynamics *212*, 167-169.

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

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

Fey, P., and Cox, E. C. (1997). Gene Trapping with GFP: the isolation of developmental mutants in the slime mould *Polysphondylium*. Current Biology *7*, 909-912.

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

Freyaldenhoven, B. S., Freyaldenhoven, M. P., Iacovoni, J. S., and Vogt, P. K. (1997). Aberrant cell growth induced by avian winged helx proteins. Cancer Research *57*, 123-129.

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

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

Fujio, K., Hu, Z., Evarts, R. P., Marsden, E. R., Niu, C. H., and Thorgeirsson, S. S. (1996). Coexpression of stem cell factor and c-kit in embryonic and adult liver. Experimental Cell Research *224*, 243-250.

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

Gardiner-Garden, M., and Frommer, M. (1987). CpG islands in vertebrate genomes. Journal of Molecular Biology *196*, 261-282.

Garrick, D., Fiering, S., Martin, D. I. K., and Whitelaw, E. (1998). Repeat-induced gene silencing in mammals. Nature Genetics *18*, 56-59.

Gasca, S., Hill, D. P., Klingensmith, J., and Rossant, J. (1995). Characterisation of a gene trap insertion into a novel gene, *cordon-bleu*, expressed in axial structures of the gastrulating mouse embryo. Developmental Genetics *17*, 141- 154.

Gerlach, C., Sakkab, D. Y., Scholzen, T., Dassler, R., Alison, M. R., and Gerdes, J. (1997). Ki-67 expression during rat liver regeneration after partial hepatectomy. Hepatology *26*, 573-578.
Gherardi, E., Gray, J., Stoker, M., Perryman, M., and Furlong, R. (1989). Purification of scatter factor, a fibroblast-derived protein that modulates epithelial interactions and movement. Proceedings of the National Academy of Science, USA *86*, 5844-5848.

Giroux, S., and Charron, J. (1998). Defective development of the embryonic liverin N-*myc*-deficient mice. Developmental Biology *195*, 16-28.

Godwin, A. R., Stadler, H. S., Nakamura, K., and Capecchi, M. R. (1998). Detectionof targeted *GFP-Hox* gene fusions during mouse embryogenesis. Proceedings of the National Academy of Science, USA *95*, 13042-13047.

Golding, M., Sarraf, C. E., Lalani, E.-N., Anilkumar, T. V., Edwards, R. J., Nagy, P., Thorgeirsson, S. S., and Alison, M. R. (1995). Oval cell differentiation into hepatocytes in the acetylaminofluorene-treated regenerating rat liver. Hepatology *22*, 1243-1253.

Gorina, S., and Pavletich, N. P. (1996). Structure of the p53 tumor suppressor bound to the ankyrin and SH3 domains of 53BP2. Science *274*, 1001-1005.

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

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

Gualdi, R., Bossard, P., Zheng, M., Hamada, Y., Coleman, J. R., and Zaret, K. S. (1996). Hepatic specification of the gut endoderm in vitro: cell signalling and transcriptional control. Genes and Development *10*, 1670- 1682.

Hadjantonakis, A. K., Gertsenstein, M., Ikawa, M., Okabe, M., and Nagy, A. (1998). Generating green flourescent mice by germline transmission of green flourescent ES cells. Mechanisms of Development *76*, 79-90.

Haffter, P., Granato, M., Brand, M., Mullins, M. C., Hammerschmidt, M., Kane, D. A., Odenthal, J., vanEeden, F., Jiang, Y. J., Heisenberg, C. P., Kelsh, R. N., FurutaniSeiki, M., Vogelsang, E., Beuchle, D., Schach, U., Fabian, C., and Nusslein-Volhard, C. (1996). The identification of genes with unique and essential functions in the development of the zebrafish, *Danio rerio*. Development *123*, 1-36.

Hammerton, R. W., Krzeminski, K. A., Mays, R. W., Ryan, T. A., Wollner, D. A., and Nelson, W. J. (1991). Mechanism for regulating cell surface distribution of Na+, K+-ATPase in polarized epithelial cells. Science *254*, 847-850.

Harrison, S. M., Dunwoodie, S. L., Arkell, R. M., Lehrach, H., and Beddington, R. S. P. (1995). Isolation of novel tissue-specific genes from cDNA libraries representing the individual tissue constituents of the gastrulating mouse embryo. Development *121*, 2479-2489.

Hartung, S., Jaenisch, R., and Breindl, M. (1986). Retrovirus insertion inactivates mouse a1(I) collagen gene by blocking initiation of transcription. Nature *320*, 365-367.

Heim, R., Prasher, D. C., and Tsien, R. Y. (1994). Wavelength mutations and posttranslational autoxidation of green flourescent protein. Proceedings of the National Academy of Science, USA *91*, 12501-12504.

Hentsch, B., Lyons, I., Li, R., Hartley, L., Lints, T. J., Adams, J. M., and Harvey, R. P. (1996). *Hlx* homeo box gene is essential for an inductive tissue interactionthat drives expansion of embryonic liver and gut. Genes and Development *10*, 70- 79.

Hermann, B. G., Labeit,S, Poustka,A King,T and Lehrach,H (1990). Cloning of the T gene required in mesoderm formation in the mouse. Nature *343*, 617-622.

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

Hicks, G. R., and Raikhel, N. V. (1995). Protein import into the nucleus: an integrated view. Annual Review of Cell and Developmental Biology *11*, 155-188.

Hilberg, F., Aguzzi, A., Howells, N., and Wagner, E. F. (1993). c-Jun is essential for normal mouse development and hepatogenesis. Nature *365*, 179- 181.

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

Hixson, D. C., Chapman, L., McBride, A., Faris, R., and Yang, L. (1997). Antigenic phenotypes common to rat oval cells, primary hepatocellular carcinomas and developing bile ducts. Carcinogenesis *18*, 1169-1175.

Hogan, B., Beddington, R., Costanini, F., and Lacy, E. (1994). *Manipulating the Mouse Embryo. A Laboratory Manual.*, 2nd Edition: Cold Spring Harbor Laboratory Press).

Hope, I. A. (1991). 'Promoter Trapping' in Caenorhabditis elegans. Development *113*, 399-408.

Hotz Vitaterna, M., King, D. P., Chang, A.-M., Kornhauser, J. M., Lowry, P. L., McDonald, J. D., Dove, W. F., Pinto, L. H., Turek, F. W., and Takahashi, J. S. (1994). Mutagenesis and mapping of a mouse gene, clock, essential for circadian behaviour. Science *264*, 719-725.

Houssaint, E. (1980). Differentiation of the mouse hepatic primordium. I. An analysis of the tissue interactions in hepatocyte differentiation. Cell Differentiation *9*, 269- 279.

Huang, E., Nocka, D. R., Beier, D., Chu, T., -Y, Buck, J., Lahm, H., -W, Wellner, D., Leder, P., and Besmer, P. (1990). The haematopoietic growth factor KL is encoded at the Steel locus and is the ligand of the c-kit receptor, the gene product of the W locus. Cell *63*, 225-233.

Iannoccone, P. M., Zhou, X., Khokha, M., Boucher, D., and Kuehn, M. R. (1992). Insertional mutation of a gene involved in growth regulation of the early mouse embryo. Developmental Dynamics *194*, 198- 208.

Isfort, R. J., Cody, D. B., Doersen, C. J., Richards, W. G., Yoder, B. K., Wilkinson, J. E., Kier, L. D., Jirtle, R. L., Isenberg, J. S., Klounig, J. E., and Woychik, R. P. (1997). The tetratricopeptide repeat containing Tg737 gene is a liver neoplasia tumor suppressor gene. Oncogene *15*, 1797-1803.

Jaenisch, R. (1988). Transgenic Animals. Science *240*, 1468-1474.

Johansson, B. M., and Wiles, M. V. (1995). Evidence for involvement of activinA and bone morphogenic protein 4 in mammalian mesoderm and hematopoietic development. Molecular and Cellular Biology *15*, 141-151.

Jones, R. O. (1970). Ultrastructural analysis of haematopoiesis in the foetal mouse. Journal of Anatomy *107*, 301-314.

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.

Jung, J., Zheng, M., Goldfarb. M., Zaret, K. S. (1999). Initiation of mammalian liver development from endoderm by fibroblast growth factors. Science *284*, 1998-2003.

Kastner, P., Messadeq, N., Mark, M., Wendling, O., Grondona, J. M., Ward, S., Ghyselinck, N., and Chambon, P. (1997). Vitamin A deficiency and mutations of RXRa, RXRb and RARa lead to early differentiation of embryonic ventricular cardiomyocytes. Development *124*, 4749-4758.

Kemphues, K. (1988). Genetic analysis of *Caenorhabditis elegans*. In Developmental Analysis of Higher Organisms, G. M. Malacinski, ed. (New York: Macmillan), pp. 193-219.

Kessel, M., and Gruss, P. (1990). Murine Developmental Control Genes. Science *249*, 374-379.

Korn, R., Schoor, M., Neuhaus, H., Henseling, U., Soininen, R., Zachgo, J., and Gossler, A. (1992). Enhancer trap integrations in mouse embryonic stem cells give rise to staining patterns in chimeric embryos with a high frequency and detect endogenous genes. Mechanisms of Development *39*, 95-109.

Kozak, M. (1996). Interpreting cDNA sequence: some insigts from studies on translation. Mammalian Genome *7*, 563-574.

Kuo, C. T., Morrisey, E. E., Anandappa, R., Sigrist, K., Lu, M. M., Parmacek, M. S., Soudais, C., and Leiden, J. M. (1997). GATA4 transcription factor is required for ventral morphogenesis and heart tube formation. Genes and Development *11*, 1048-60.

Kwok, S., and Higuchi, R. (1989). Avoiding false positives with PCR. Nature *339*, 237- 238.

Laemmli, U. K. (1970). Cleavage of the structural protein during the assembly of the head of bacteriophage T4. Nature *227*, 680-685.

Lai, E., Clark, K. L., Burley, S. K., and Darnell, J. E., Jr. (1993). Hepatocyte nuclear factor3/forkhead or "winged helix" proteins: A family of transcription factors of diverse biological function. Proceedings of the National Academy of Science, USA *90*, 10421-10423.

Lambert, S., Yu, H., Prchal, J. T., Lawler, J., Ruff, P., Speicher, D., Cheung, M. C., Kan, Y. W., and Palek, J. (1990). cDNA sequence for human erythrocyte ankyrin. Proceedings of the National Academy of Science, USA *87*, 1730- 1734.

Le Douarin, N. M. (1975). An experimental analysis of liver development. Medical Biology *53*, 427- 455.

Lee, S. W., Tomasetto, C., and Sager, R. (1991). Positive selection of candidate tumor-suppressor genes by subtractive hybridisation. Proceedings of the National Academy of Science, USA *88*, 2825-2829.

Lee, W., Mitchell, P., and Tijan, R. (1987). Purified transcription factor AP-1interacts with TPA-inducible enhancer elements. Cell *49*, 741-752.

Leffert, H., Moran, T., Sell, S., Skelly, H., Ibsen, K., Mueller, M., and Arias, I. (1978). Growth state-dependent phenotypes of adult hepatocytes in primary culture. Proceedings of the National Academy of Science, USA *75*, 1834-1838.

Liang, P., and Pardee, A. B. (1992). Differential display of eukaryotic messenger RNA by means of the polymerase chain reaction. Science *257*, 967-971.

Lints, T. J., Hartley, L., Parsons, L. M., and Harvey, R. P. (1996). Mesoderm-specific expression of the divergent homeobox gene Hlx during murine embryogenesis. Developmental Dynamics *205*, 457-470.

Lux, S. E., John, K. M., and Bennet, V. (1990). Analysis of cDNA for human erythrocyte ankyrin indicates a repeated structure with homology to tissue-differentiation and cell cycle control proteins. Nature *344*, 36-42.

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.

Marshall, H., Nonchev, S., Sham, M. H., Muchamore, I., Lumsden, A., and Krumlauf, R. (1992). Retinoic acid alters hindbrain *Hox* code and induces transformation of rhombomeres 2/3 into a 4/5 identity. Nature *360*, 737-741.

Martin, G. M. (1981). Isolation of a pluripotent cell line from early mouse embryos cultured in medium conditioned by teratocarcinoma stem cells. Proceedings of the National Academy of Science, USA *78*, 7634-7638.

McClive, P., Pall, G., Newton, K., Lee, M., Mullins, J., and Forrester, L. (1998). Gene trap integrations expressed in the developing heart: insertion site affects splicing of the PT1-ATG vector. Developmental Dynamics *212*, 267-276.

McMahon, A. P. (1992). The Wnt family of developmental regulators. Trends in Genetics *8*, 236-242.

Medvinsky, A. L., Samoylina, N. L., Muller, A. M., and Dzierzak, E. A. (1993). An early embryonic pre-liver intra-embryonic source of CFU-S in the developing mouse. Nature *364*, 64- 66.

Meehan, R. R., Barlow, D. P., Hill, R. E., Hogan, B. L. M., and Hastie, N. D. (1984). Pattern of serum protein gene expression in mouse visceral yolk sac and foetal liver. EMBO Journal *3*, 1881-1885.

Michaely, P., and Bennett, V. (1992). The ANK repeat: a ubiquitous motif involved in macromolecular recognition. Trends in Cell Biology *2*, 127-129.

Michaely, P., and Bennett, V. (1993). The membrane-binding domain of ankyrin contains four independently folded subdomains, each comprised of six ankyrin repeats. The Journal of Biological Chemistry *268*, 22703-22709.

Miklos, G. L. G., and Rubin, G. M. (1996). The role of the genome project in determining gene function: Insights from model organisms. Cell *86*, 521-529.

Molkentin, J. D., Lin, Q., Duncan, S. A., and Olson, E. N. (1997). Requirement of the transcription factor GATA4 for heart tube formation and ventral morphogenesis. Genes and Development *11*, 1061- 1072.

Monaghan, A. P., Kaestner, K. H., Grau, E., and Schutz, G. (1993). Postimplantation expression patterns indicate a role for the mouse *forkhead*/HNF-3a, b and c genes in determination of the definitive endoderm, chordamesoderm and neuroectoderm. Development *119*, 567-578.

Montesano, R., Schaller, G., and Orci, L. (1991). Induction of epithelial tubular morphogenesis *in vitro* by fibroblast derived soluble factors. Cell *66*, 697-711.

Moore, M. R., Drinkwater, N. R., Miller, E. C., Miller, J. A., and Pilot, H. C. (1981). Quantitive analysis of the time-dependant development of glucose-6-phosphatase-deficient foci in the livers of mice treated neonatally with diethylnitrosamine. Cancer Research *41*, 1585- 1593.

Mooslehner, K., Karls, U., and Harbers, K. (1990). Retroviral integration sites in transgenic Mov mice frequently map in the vicinity of transcribed DNA regions. Journal of Virology *64*, 3056-3058.

Morgan, D., Turnpenny, L., Goodship, J., Dai, W., Majumder, K., Matthews, L., Gardner, A., Schuster, G., Vien, L., Harrison, W., Elder, F. F. B., Penman-Splitt, M., Overbeek, P., and Strachan, T. (1998). Inversin, a novel gene in the vertebrate left-right axis pathway, is partially deleted in the *inv* mouse. Nature Genetics *20*, 149-156.

Motoyama, J., Kitajima, K., Kojima, M., Kondo, S., and Takeuchi, T. (1997). Organogenesis of the liver, thymus and spleen is affected in *jumonji* mutant mice. Mechanisms of Development *66*, 27-37.

Mountford, P. S., and Smith, A. G. (1995). Internal ribosome entry sites and dicistronic RNAs in mammalian transgenesis. Trends in Genetics *11*, 179-184.

Muth, K., Bruyns, R., Thorey, I. S., and von Melcher, H. (1998). Disruption of genes regulated during hematopoietic differentiation of mouse embryonic stem cells. Developmental Dynamics *212*, 277-283.

Nagy, A., Gocza, E., Diaz, E. M., Prideaux, V. R., Ivanyi, E., Markkula, M., and Rossant, J. (1990). Embryonic stem cells alone are able to support fetal development in the mouse. Development *110*, 815-821.

Nakamura, T., Nishizawa, T., Hagiya, M., Seki, T., Shimonishi, M., Sugimura, A., Tashiro, K., and Shimizu, S. (1989). Molecular cloning and expression of human hepatocyte growth factor. Nature *342*, 440-443.

Narita, N., Bielinska, M., and Wilson, D. B. (1997a). Cardiomyocyte differentiation by GATA-4 deficient embryonic stem cells. Development *122*, 3764.

Narita, N., Bielinska, M., and Wilson, D. B. (1997b). Wild type visceral endoderm abrogates the ventral developmental defects associated with GATA-4 deficiency in the mouse. Developmental Biology *189*, 270- 274.

Nelson, W. J., and Veshnock, P. J. (1987). Ankyrin binding to (Na+/K+)ATPase and the implication for the organisation of the of membrane domains in polarized cells. Nature *328*, 533-536.

Neuhaus, H., Bettenhausen, B., Bilinski, P., Simon-Chazottes, D., Guénet, J.-L., and Gossler, A. (1994). Etl2, A novel putative type-I cytokine receptor expressed during mouse embryogenesis at high levels in skin and cells with skeletogenic potential. Developmental Biology *166*, 531-542.

Niwa, H., Araki, K., Kimura, S., Taniguchi, S., Wakasugi, S., and Yamamura, K. (1993). An efficient gene-trap method using poly-a trap vectors and characterization of gene-trap events. Journal of Biochemistry *113*, 343-349.

Nüsslein-Volhard, C., Wiescaus, E., and Kluding, H. (1984). Mutations affecting the pattern of the larval cuticle in *Drosophila melanogaster* I. Zygotic loci on the second chromosome. Roux's Archives of Developmental Biology *193*, 267-282.

O'Kane, C., and Gehring, W. (1987). Detection in situ of genomic regulatory elements in *Drosophila*. Proceedings of the National Academy of Science, USA *84*, 9123-9127.

Otto, E., Kunimoto, M., McLaughlin, T., and Bennett, V. (1991). Isolation and characterisation of cDNAs encoding human brain ankyrins reveal a family of alternatively spliced genes. The Journal of Cell Biology *114*, 241-253.

Overturf, K., Al-Dhalimy, M., Ou, C. N., Finegold, M., and Grompe, M. (1997). Serial transplantation reveals the stem-like regenerative potential of adult mouse hepatocytes. American Journal of Anatomy *151*, 1273-1280.

Peters, L. H., John, K. M., Lu, F. M., Eicher, E. M., Higgins, A., Yialamas, M., Turtzo, L. C., Otsuka, A. J., and Lux, S. E. (1995). Ank3 (epithelial ankyrin), a widely distributede new member of the ankyrin gene family and the major ankyrin in kidney, is expressed in alternatively spliced forms, including forms that lack the repeat domain. The Journal of Cell Biology *130*, 313-330.

Pires-DaSilva, A., and Gruss, P. (1998). Gene trap insertion into a novel gene expressed during mouse limb development. Developmental Dynamics *212*, 318-325.

Porter-Jordan, K., and Garrett, C. T. (1990). Source of contamination in polymerase chain reaction assay. Lancet *335*, 1220.

Pruitt, S. C. (1992). Expression of Pax-3- and neuroectoderm-inducing activities during differentiation of P19 embryonal carcinoma cells. Development *116*, 573-583.

Rabbits, P., Impey, H., Heppel-Parton, A., Langford, C., Tease, C., Lowe, N., Bailey, D., Ferguson-Smith, M., and Carter, N. (1995). Chromosome specific paints from a high resolution flow karyotype of the mouse. Nature Genetics *9*, 369-375.

Ransone, L. J., and Verma, I. M. (1990). Nuclear proto-oncogenes *fos* and *jun*. Annual Review of Cell Biology *6*, 539-557.

Rehorn, K.-P., Thelen, H., Michelson, A. M., and Reuter, R. (1996). A molecular aspect of hematopoiesis and endoderm development common to vertebrates and *Drosophila*. Development *122*, 4023-4031.

Reith, A. D., Rottapel, R., Giddens, E., Brady, C., Forrester, L., and Bernstein, A. (1990). W mutant mice with mild or severe developmental defects contain distinct point mutations in the kinase domain of the c-kit receptor. Genes and Development *4*, 390-400.

Rijkers, T., Peetz, A., and Ruther, U. (1994). Insertional mutagenesis in transgenic mice. Transgenic Research *3*, 203- 215.

Rinchik, E. M. (1991). Chemical mutagenesis and fine-structure functional analysis of the mouse genome. Trends in Genetics *7*, 15-21.

Robertson, E., Bradley, A., Kuehn, M., and Evans, M. (1986). Germ-line transmission of genes introduced into cultured pluripotential cells by a retroviral vector. Nature *309*, 255-256.

Rodrigues, G. A., Park, M., and Schlessinger, J. (1997). Activation of the JNK pathway is essential for transformation by the Met oncogene. EMBO Journal *16*, 2634-2645.

Rogers, D. C., Fisher, E. M., Brown, S. D., Peters, J., Hunter, A. J., and Martin, J. E. (1997). Behavioral and functional analysis of mouse phenotype: SHIRPA, a protocol for comprehensive phenotype assessment. Mammalian Genome *8*, 711-713.

Rogers, S., Wells, R., and Rechsteiner, M. (1986). Amino acid sequences common to rapidly degraded proteins: the PEST hypothesis. Science *234*, 364-368.

Rohdewohld, H., Weiher, H., Reik, W., Jaenisch, R., and Breindl, M. (1987). Retrovirus integration and chromatin structure: Moloney murine leukemia proviral integration sites map near DNase I-hypersensitive sites. Journal of Virology *61*, 336-343.

Rohwedel, J., Maltsev, V., Bober, E., Arnold, H. H., Hescheler, J., and Wobus, A. M. (1994). Muscle cell differentiation of embrynic stem cells reflects myogenesis *in vivo*: developmentally regulated expression of myogenic genes and functional expression of ionic currents. Developmental Biology *164*, 87-101.

Russ, A. P., Friedel, C., Ballas, K., Kalina, U., Zahn, D., Strebhardt, K., and vonMelchner, H. (1996). Identification of genes induced by factor deprivation in hematopoietic cells undergoing apoptosis using gene-trap mutagenesis and site-specific recombination. Proceedings of the National Academy of Science, USA *93*, 15279-15284.

Sam, M., Wurst, W., Kluppel, M., Jin, O., Heng, H., and Bernstein, A. (1998). *Aquarius*, a novel gene isolated by gene trapping with an RNA-dependent RNA polymerase motif. Developmental Dynamics *212*, 304-317.

Sambrook, J., Fritsch, E. F., and Maniatis, T. (1989). Molecular cloning, a laboratory manual (Cold Spring Harbor: CSH laboratory press).

Sanford, L. P., Ormsby, I., Gittenberger-de Groot, A. C., Sariola, H., Friedman, R., Boivin G. P., Cardell E. L., Doetschman, T., (1997). TGFbeta2 knockout mice have multiple developmental defects that are non-overlapping with other TGFbeta knockout phenotypes. Development *124*, 2659-2670

Sanger, F., Niklen, S., and Coulson, A. R. (1977). DNA sequencing with chain-terminating inhibitors. Proceeding of the National Academy of Science, USA *74*, 5463-5467.

Sasaki, H., and Hogan, B. L. M. (1993). Differential expression of multiple fork head related genes during gastrulation and axial pattern formation in the mouse embryo. Development *118*, 47-59.

Scherdin, U., Rhodes, K., and Breindl, M. (1990). Transcriptionally Active Genome Regions are preferred targets for retrovirus integration. Journal of Virology *64*, 907-912.

Scherer, C. A., Chen, J., Nachabeh, A., Hopkins, N., and Ruley, H. E. (1996). Transcriptional specificity of the pluripotent embryonic stem cell. Cell Growth and Differentiation *7*, 1393-1401.

Schmidt, C., Bladt, F., Goedecke, S., Brinkman, V., Zschiesche, W., Sharpe, M., Gherardi, E., and Birchmeier, C. (1995). Scatter factor/hepatocyte growth factor is esential for liver development. Nature *373*, 699- 702.

Schuster-Gossler, K., Simonchazottes, D., Guenet, J. L., Zachgo, J., Gossler, A. (1996). Gtl2(lacZ) an insertional mutation on mouse chromosome-12 with parental origin-dependent phenotype. Mammalian Genome *7*, 20- 24.

Schuster-Gossler, K., Bilinski, P., Sado, T., Ferguson-Smith, A., and Gossler, A. (1998). The mouse Gtl2 gene is differentially expressed during embryonic development, encodes multple alternatively spliced transcripts, and may act as an RNA. Developmental Dynamics *212*, 214-228.

Serafini, T., Colamarino, S. A., Leonardo, E. D., Wang, H., Beddington, R., Skarnes, W. C., and Tessier-Lavigne, M. (1996). *Netrin-1* is required for commisural axon guidance in the developing vertebrate nervous system. Cell *87*, 1001-1014.

Severn, C. B. (1972). A morphological study of the development of the human liver. American Journal of Anatomy *133*, 85-108.

Shedlovsky, A., King, T. R., and Dove, W. F. (1988). Saturation germ line mutagenesis of the murine *t* region including a lethal allele at the *quaking* locus. Proceedings of the National Academy of Science, USA *85*, 180-184.

Shedlovsky, A., McDonald, J. D., Symula, D., and Dove, W. F. (1993). Mouse models of human phenylketonuria. Genetics *134*, 1205-1210.

Shih, C.-C., Stoye, J. P., and Coffin, J. M. (1988). Highly preferred targets for retroviral integration. Cell *53*, 531-537.

Shiojiri, N. (1981). Enzymo- and immunocytochemical analyses of the differentiation of liver cells in the prenatal mouse. Journal of Embryology and Experimental Morphology *62*, 139-152.

Sidow, A., Bulotsky, M. S., Kerrebrock, A. W., Bronson, R. T., Daly, M. J., Reeve, M. P., Hawkins, T. L., Birren, B. W., Jaenisch, R., and Lander, E. S. (1997). Serrrate2 is disrupted in the mouse limb-development mutant syndactylism. Nature *389*, 722-725.

Simeone, A., Acampora, D., Arcioni, L., Andrews, P. W., Boncinelli, E., and Mavilio, F. (1990). Sequential activation of *HOX2* homeobox genes by retinoic acid in human embryonal carcinoma cells. Nature *346*, 763-766.

Skarnes, W. C. (1990). Entrapment vectors : A new tool for mammalian genetics. Biotechnology *8*, 827-831.

Skarnes, W. C., Auerbach, B. A., and Joyner, A. L. (1992). A gene trap approach in mouse embryonic stem cells: the *lacZ* reporter is activated by splicing, reflects endogenous gene expression and is mutagenic in mice. Genes and 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 Science, USA *92*, 6592- 6596.

Sladek, F. M., Zhong, W., Lai, E., and Darnell, J. E., Jr. (1990). Liver-enriched transcription factor HNF-4 is a novel member of the steroid hormone receptor superfamily. Genes and Development *4*, 2353-2365.

Smith, A. G. (1991). Culture and differentiation of embryonic stem cells. Journal of Tissue Culture Methods *13*, 89-94.

Smith, D., Wohlgemuth, J., Calvi, B. R., Franklin, I., and Gelbart, W. M. (1993). *hobo* Enhancer Trapping Mutagenesis in Drosophila Reveals an Insertion Specificity Differenet from P elements. Genetics *135*, 1063-1076.

Soininen, R., Schor, M., Henseling, U., Tepe, C., Kisters-Woike, B., Rossant, J., and Gossler, A. (1992). The mouse enhancer trap locus 1 (Etl-1): a novel mammalian gene related to Drosophila and yeast transcriptional regulator genes. Mechanisms of Development *39*, 111-123.

Sonnenberg, E., Meyer, D., Michael Weidner, K., and Birchmeier, C. (1993). Scatter factor/hepatocyte growth factor and its receptor, c-met tyrosine kinase, can mediate a signal exchange between mesenchyme and epithelia during mouse development. Journal of Cell Biology *123*, 223-235.

Spradling, A. C., Stern, D. M., Kiss, I., Roote, J., Laverty, T., and Rubin, G. M. (1995). gene disruptions using P transposable elements: An integral component of the *Drosophila* genome project. Proceedings of the National Academy of Science, USA *92*, 10824-10830.

Stoker, M., Gherardi, E., Perryman, M., and Gray, J. (1987). Scatter factor is a fibroblast-derived modulator of epithelial cell mobility. Nature *327*, 239-242.

Stoykova, A., Chowdhury, K., Bonaldo, P., Torres, M., and Gruss, P. (1998). Gene trap expression and mutational analysis for genes involved in the development of the mammalian nervous system. Developmental Dynamics *212*, 198-213.

Takeuchi, T., Yamazaki, Y., Katoh-Fukui, Y., Tsuchiya, R., Kondo, S., Motoyama, J., and Higashinakagawa, T. (1995). Gene trap capture of a novel mouse gene, *jumonji*, required for neural tube formation. Genes and Development *9*, 1211- 1222.

Thomas, P. Q., Brown, A., and Beddington, R. S. P. (1998a). *Hex*: a homeobox gene revealing peri-implantation asymmetry in the mouse embryo and an early transient marker of endothelial cell precursors. Development *125*, 85-94.

Thomas, T., Voss, A. K., and Gruss, P. (1998b). Distribution of a murine protein tyrosine phosphatase BL-b-galactosidase fusion protein suggests a role in neurite outgrowth. Developmental Dynamics *212*, 250-257.

Thompson, S., Clarke, A. R., Pow, A. M., Hooper, M. L., and Melton, D. W. (1989). Germ line transmission and expression of a corrected HPRT gene produced by gene targeting in embryonic stem cells. Cell *56*, 313-321.

Thorey, I. S., Muth, K., Russ, A. P., Otte, J., Reffelmann, A., and von Melcher, H. (1998). Selective disruption of genes transiently induced in differentiating mouse embryonic stem cells by using gene trap mutagenesis and site specific recombination. Molecular and Cellular Biology *18*, 3081-3088.

Torres, M., Stoykova, A., Huber, O., Chowdhury, K., Bonaldo, P., Mansouri, A., Butz, S., Kemler, R., and Gruss, P. (1997). An *a-E-catenin* gene trap mutation defines its function in preimplantation development. Proceedings of the National Academy of Science, USA *94*, 901-906.

Townley, D. J., Avery, B. J., Rosen, B., and Skarnes, W. C. (1997). Rapid sequence analysis of gene trap integrations to generate a resource of insertional mutations in mice. Genome Research *7*, 293- 298.

Tyner, A. L., Godbout, R., Compton, R. S., and Tilghman, S. M. (1990). The ontogeny of a-fetoprotein gene expression in the mouse gastrointestinal tract. The Journal of Cell Biology *110*, 915-927.

Uriel, J., Bouillon, D., Aussel, C., and Dupiers, M. (1976). Alpha-fetoprotein: The major high-affinity estrogen binder in rat uterine cytosols. Proceedings of the National Academy of Science, USA *73*, 1452-1456.

Venkataramani, R., Swaminathan, K., and Marmorstein, R. (1998). Crystal structure of the CDK4/6 inhibitory protein $p18^{INK4c}$ provides insights into ankyrin-like repeat structure/function and tumor-deived $p16^{INK4}$ mutation. Nature Structural Biology *5*, 74-81.

Vijaya, S., Steffen, D. L., and Robinson, H. L. (1986). Acceptor sites for retroviral integrations map near DNase I-hypersensitive sites in chromatin. Journal of Virology *60*, 683-692.

Voit, R., Schnapp, A., Kuhn, A., Rosenbauer, H., Hirschmann, P., Stunnenberg, H. G., and Grummt, I. (1992). The nucleolar transcription factor mUBF is phosphorylated by casein kinase II in the C-terminal hyperacidic tail which is essential for transactivation. EMBO Journal *11*, 2211-2218.

von Melchner, H., DeGregori, J. V., Rayburn, H., Reddy, S., Friedel, C., and Ruley, H. E. (1992). Selective disruption of genes expressed in totipotent embryonal stem cells. Genes and Development *6*, 919- 927.

von Melchner, H., Reddy, S., and Ruley, H. E. (1990). Isolation of cellular promoters by using a retrovirus promoter trap. Proceedings of the National Academy of Science, USA *87*, 3733- 3737.

Voss, A. K., Thomas, T., and Gruss, P. (1998b). Compensation for a gene trap mutation in the murine microtubule-associated protein 4 locus by alternative polyadenylation and alternative splicing. Developmental Dynamics *212*, 258-266.

Voss, A. K., Thomas, T., and Gruss, P. (1998a). Efficiency assessment of the gene trap approach. Developmental Dynamics *212*, 171-180.

Weigel, D., Jurgens, G., Kuttner, F., Seifert, E., and Jackle, H. (1989). The homeotic gene fork head encodes a nuclear protein and is expressed in the terminal regions of the *Drosophila* embryo. Cell *57*, 645-658.

Welsh, J., Chada, K., Dalal, S. S., Cheng, R., Ralph, D., and McClelland, M. (1992). Arbitrarily primed PCR fingerprinting of RNA. Nucleic Acids Research *20*, 4965-4970.

Wertz, K., and Fuchtbauer, E.-M. (1998). *Dmd* $^{mdx\text{-}bgeo}$: A new allele for the mouse dystrophin gene. Developmental Dynamics *212*, 229-241.

Wilson, C., Pearson, R. K., Bellen, H. J., O'Kane, C., Grossniklaus, U., and Gehring, W. J. (1989). P-element-mediated enhancer detection: an efficient method for isolating and characterising developmentally regulated genes in *Drosophila*. Genes and Development *3*, 1301-1313.

Winship, P. R. (1989). An improved method for directly sequencing PCR amplified material using dimethyl sulfoxide. Nucleic Acids Research *17*, 1266.

Withers-Ward, E. S., Kitamura, Y., Barnes, J. P., and Coffin, J. M. (1994). Distribution of targets for avain retrovirus DNA integration *in vivo*. Genes and Development *8*, 1473-1487.

Wolberg, C. (1998). Combinatorial transcription factors. Current Opinion in Genetics and Development *8*, 552-559.

Wolffe, A. P. (1998). When more is less. Nature Genetics *18*, 5-6.

Xiong, J.-W., Battaglino, R., Leahy, A., and Stuhlmann, H. (1998). Large-scale screening for developmental genes in embryonic stem cells and embryoid bodies using retroviral entrapment vectors. Developmental Dynamics *212*, 181-197.

Yamaguchi, T. P., Harpal, K., Henkemeyer, M., and Rossant, J. (1994). fgfr-1 is required for embryonic growth and mesodermal patterning during mouse gastrulation. Genes and Development *8*, 3032-3044.

Yenofsky, R. L., Fine, M., and Pellow, J. W. (1990). A mutant neomycin phosphotransferase-II gene reduces the resistance of transformants to antibiotic selection pressure. Proc Natl Acad Sci *87*, 3435-3439.

Yoshida, M., Yagi, T., Furuta, Y., Takayanagi, K., Kominami, R., Takeda, N., Tokunaga, T., Chiba, J., Ikawa, Y., and Aizawa, S. (1995). A new strategy of gene trapping in ES cells using 3' RACE. Transgenic Research *4*, 277-287.

Zambrowicz, B. P., Friedrich, G. A., Buxton, E. C., Lilleberg, S. L., Person, C., and Sands, A. T. (1998). Disruption and sequence identification of 2,000 genes in mouse embryonic stem cells. Nature *392*, 608- 611.

Zambrowicz, B. P., Imamoto, A., Fiering, S., Herzenberg, L. A., Kerr, W. G., and Soriano, P. (1997). Disruption of overlapping transcripts in the ROSAβgeo 26 gene trap strain leads to widespread expression of b-galactosidase in mouse embryos and hematopoietic cells. Proceedings of the National Academy of Science, USA *94*, 3789-3794.

Zaret, K. (1998). Early liver differentiation: genetic potentiation and multilelvel growth control. Current Opinions in Genetics and Development *8*, 526-531.

Zaret, K. S. (1996). Molecular genetics of early liver development. Annual Review of Physiology *58*, 231- 251.

Zarnegar, R., and Michalopoulos, G. (1989). Purification and biological characterisation of human hepatopoietin A, a polypeptide growth factor for hepatocytes. Cancer Research *49*, 3314-3320.

Zhou, X., Sasaki, H., Lowe, L., Hogan, B. L. M., and Kuehn, M. R. (1993). *Nodal* is a novel TGF-b-like gene expressed in the mouse node during gastrulation. Nature *361*, 543- 547.

Zhu, J., Hill, R. J., Heid, P. J., Fukuyama, M., Sugimoto, A., Priess, J. R., and Rothman, J. H. (1997). *End-1* encodes an apparent GATA factor that specifies the endoderm precursor in *Caenorhabditis elegans* embryos. Genes and Development *11*, 2883-2896.