Retroelements as controlling elements in mammals

Author:
Thomson, Gabrielle Anne

Publication Date:
2006

DOI:
https://doi.org/10.26190/unsworks/16081

RETROELEMENTS AS CONTROLLING ELEMENTS IN MAMMALS

Gabrielle Thomson

BSc (Hon)

A thesis submitted for the degree of Doctor of Philosophy

Victor Chang Cardiac Research Institute, St Vincents Hospital

and the

School of Biotechnology and Biomolecular Sciences

April 2006

PLEASE TYPE

THE UNIVERSITY OF NEW SOUTH WALES
Thesis/Dissertation Sheet

Surname or Family name: Thomson

First name: Gabrielle Other names: Anne

Abbreviation for degree as given in the University Calendar: PhD

School: School of Biotechnology and Biomolecular Sciences Faculty: Science

Title: Retroelements as controlling elements in mammals

Abstract 350 words maximim: (PLEASE TYPE)

Retroelements are genomic parasites which make up ~42% of the human genome and 38% of the mouse genome. Most are degenerate, but a large number have relatively intact promoter elements, suggesting that they are capable of transcription. Transcriptionally active retroelements can perturb normal transcription units in their vicinity through a variety of mechanisms, leading to phenotypic effects and in some cases disease. This phenomenon of transcriptional interference has been observed in organisms as diverse as maize, Drosophila, and the mouse. We analysed the extent of retroelement transcription in normal and diseased tissues, by searching the mouse and human EST databases for transcripts originating in retroelement promoters, and found a large number of transcripts from LINEs, SINEs and ERVs. Retroelement transcripts were found to be initiated in both sense and antisense orientations, and to be equally as common in normal and diseased tissue. Several of these transcripts were chimeric, appearing to initiate in a retroelements and reading through to cellular genes, suggestive of transcriptional interference. We have used transposon display to identify and recover retroelement transcripts in the mouse. Transcripts initiated in LINE, SINE and ERV promoters are numerous, and many are chimeric with cellular genes. Although the numbers of recovered chimeric transcripts are too large to permit rigorous analysis of more than a small proportion, some of those we have studied further appear to be authentic transcripts that may represent interference with the canonical promoters of the genes in question. Our results suggest that transcriptional interference by retroelements may be a relatively common occurrence in mammals.

Declaration relating to disposition of project thesis/dissertation

I hereby grant to the University of New South Wales or its agents the right to archive and to make available my thesis or dissertation in whole or in part in the University libraries in all forms of media, now or here after known, subject to the provisions of the Copyright Act 1968. I retain all property rights, such as patent rights. I also retain the right to use in future works (such as articles or books) all or part of this thesis or dissertation.

I also authorise University Microfilms to use the 350 word abstract of my thesis in Dissertation Abstracts International (this is applicable to doctoral theses only).

Signature: G Thomson Witness: D.T. Humphreys Date: 20/12/06

The University recognises that there may be exceptional circumstances requiring restrictions on copying or conditions on use. Requests for restriction for a period of up to 2 years must be made in writing to the Registrar. Requests for a longer period of restriction may be considered in exceptional circumstances if accompanied by a letter of support from the Supervisor or Head of School. Such requests must be submitted with the thesis/dissertation.

FOR OFFICE USE ONLY Date of Completion of requirements for the Award:

Registrar and Deputy Principal

THIS SHEET IS TO BE GLUED TO THE INSIDE FRONT COVER OF THE THESIS

ORIGINALITY STATEMENT

'I hereby declare that this submission is my own work and to the best of my knowledge it contains no materials previously published or written by another person, or substantial proportions of material which have been accepted for the award of any other degree or diploma at UNSW or any other educational institution, except where due acknowledgement is made in the thesis. Any contribution made to the research by others, with whom I have worked at UNSW or elsewhere, is explicitly acknowledged in the thesis. I also declare that the intellectual content of this thesis is the product of my own work, except to the extent that assistance from others in the project's design and conception or in style, presentation and linguistic expression is acknowledged.'

Signed .....

COPYRIGHT STATEMENT

'I hereby grant the University of New South Wales or its agents the right to archive and to make available my thesis or dissertation in whole or part in the University libraries in all forms of media, now or here after known, subject to the provisions of the Copyright Act 1968. I retain all proprietary rights, such as patent rights. I also retain the right to use in future works (such as articles or books) all or part of this thesis or dissertation.
I also authorise University Microfilms to use the 350 word abstract of my thesis in Dissertation Abstract International (this is applicable to doctoral theses only).
I have either used no substantial portions of copyright material in my thesis or I have obtained permission to use copyright material; where permission has not been granted I have applied/will apply for a partial restriction of the digital copy of my thesis or dissertation.'

Signed .....G. Thomas.....

Date .....20/12/06.....

AUTHENTICITY STATEMENT

'I certify that the Library deposit digital copy is a direct equivalent of the final officially approved version of my thesis. No emendation of content has occurred and if there are any minor variations in formatting, they are the result of the conversion to digital format.'

Signed .....G. Thomas.....

Date .....20/12/06.....

In Memory Of Ian

ABSTRACT

Retroelements are genomic parasites which make up ~42% of the human genome and 38% of the mouse genome. Most are degenerate, but a large number have relatively intact promoter elements, suggesting that they are capable of transcription. Transcriptionally active retroelements can perturb normal transcription units in their vicinity through a variety of mechanisms, leading to phenotypic effects and in some cases disease. This phenomenon of transcriptional interference has been observed in organisms as diverse as maize, *Drosophila*, and the mouse. We analysed the extent of retroelement transcription in normal and diseased tissues, by searching the mouse and human EST databases for transcripts originating in retroelement promoters, and found a large number of transcripts from LINEs, SINEs and ERVs. Retroelement transcripts were found to be initiated in both sense and antisense orientations, and to be equally as common in normal and diseased tissue. Several of these transcripts were chimeric, appearing to initiate in retroelements and reading through to cellular genes, suggestive of transcriptional interference. We have used transposon display to identify and recover retroelement transcripts in the mouse. Transcripts initiated in LINE, SINE and ERV promoters are numerous, and many are chimeric with cellular genes. Although the numbers of recovered chimeric transcripts are too large to permit rigorous analysis of more than a small proportion, some of those we have studied further appear to be authentic transcripts that may represent interference with the canonical promoters of the genes in question. Our results suggest that transcriptional interference by retroelements may be a relatively common occurrence in mammals.

Acknowledgements

After 5 years of hard work this thesis was nearly abandoned but with the support of some wonderful people, here it is finished.

Firstly I would like to thank my supervisor David Martin for all, his time, patience and input. I have learnt some valuable lessons from him and without him this thesis would never have been started let alone finished. His kindness and support over the past 8 months in particular has been invaluable.

Thankyou also to my internal supervisor Ian Dawes for his sound advice, support and friendship.

A huge thankyou goes to Bryony and Dave H. whose kindness, support and gentle pushing helped me not only to finish but helped me survive through the past 8 months.

Thankyou to the various members of the Martin Lab who have contributed in valuable discussions and who have kept me sane – Jen (who suffered through my PhD with me), Wendy, Lachlan and Cath - I couldn't have made it without you guys! Thanks also to all my other Victor Chang buddies in particular Gill, Angela, Rob, Jost and Gavin who put up with my constant talking! The Victor Chang really supported me throughout my PhD and a big thankyou goes to everyone who was a part of that.

True friendship is the most valuable thing in life and without Bek, I don't know where I would be. She held me together when my world fell apart and helped me rebuild it again. Cathie, Jane, Freya, Ian C, Aaron and Barbara, have also helped me stand when I've stumbled and to them I am eternally grateful. Thanks also to my other wonderful friends who are too numerous to mention but who are equally appreciated.

Family support is ever enduring and a big thankyou goes to the Thomson family for being such wonderful people who understand the true meaning of family.

Thankyou to my parents, brother and sister in law who have been a constant source of inspiration and support and who have held me together through the rough times. Without you guys I couldn't have made it.

Finally I'm thankful to have had the little time I did with my husband – he gave me strength and love and was always there for me. I just wish he could have been here to see the end.

This thesis was finished for him

TABLE OF CONTENTS

TABLE OF FIGURES

TABLE OF TABLES

Abbreviations

A^{Fu}	Axin fused
A^{vy}	Agouti viable yellow
ADAR	adenosine deaminase that acts on RNA
AFLP	Amplified fragment length polymorphism
cDNA	Complementary DNA
dbEST	EST database
dsRNA	Double stranded RNA
DEPC	Diethylpyrocarbonate
DNA	Deoxyribonucleic acid
DNase	Deoxyribonuclease
Dnmt	DNA methyltransferase
ERV	Endogenous retrovirus
EST	Expressed sequence tag
ETn	Early transposon
γ-IRE	Interferon gamma responsive elements
GLN	Murine ERV with glutamine tRNA primer
HERV	Human endogenous retrovirus
IAP	Intracisternal A particle
IFN-γ	Interferon gamma
LINE	Long interspersed repeat
LTR	Long terminal repeat
MaLR	Mammalian LTR transposon
meC	Methyl cytosine

Merv	Mouse endogenous retrovirus
MHC	Major histocompatibility complex
miRNA	Micro RNA
MMTV	Mouse mammary tumour virus
mRNA	Messenger RNA
MT	Mouse transcript retroelement
MuLV	Murine leukaemia virus
MuRRS	Murine retrovirus-related sequence
MuRVY	Murine repeated virus on the Y chromosome
ORF	Open reading frame
PCR	Polymerase chain reaction
PGC	Primordial germ cell
RACE	Rapid amplification of cDNA ends
RNA	Ribonucleic acid
RNAi	RNA interference
RNase	Ribonuclease
RNH	Ribonuclease H
RT	Reverse transcriptase
SAg	Superantigen
SHR	Spontaneously hypertensive rats
SINE	Short interspersed repeat
siRNA	Silencing RNA
tRNA	Transfer RNA
UTR	Untranslated region
VL30	Virus-like 30 S

CHAPTER 1 INTRODUCTION

1.1 Retroelements and the mammalian genome

Retroelements, which have often been called "junk" DNA, make up approximately 42% of the human genome and approximately 37% of the mouse genome. This is a much higher proportion of the genome than is occupied by protein encoding genes and their associated regulatory sequences (~ 5%). Yet it is commonly believed that retroelements have no biological significance: they are assumed to play no role in phenotypic differences.

This thesis is based on the idea that retroelements alter the transcriptional regulation of genes around them, leading to phenotypic changes and in some cases disease (Whitelaw and Martin 2001). This supposition stems from demonstrations that retroelements in mice are capable of just such effects (Duhl *et al.* 1994; Vasicek *et al.* 1997).

It is commonly believed that repression of retroelements, in the form of methylation (Yoder *et al.* 1997) and possibly even RNA interference (Svoboda *et al.* 2004; Soifer *et al.* 2005), is employed to silence transcription of the vast majority of elements (if not all of them). It is also widely believed that retroelements are with very few exceptions degenerate and essentially inert, so that any effects they may have are passive. Although there is a clear basis for these beliefs, there is also a large amount of evidence to suggest that retroelements sometimes escape silencing and are active in somatic cells. The large number of retroelements in the mammalian genome, combined with the ability of some of these elements to transcribe themselves, lends support to the idea that these elements may actively contribute to phenotypic variation and even disease risk .

1.1.1 What is a retroelement?

Retroelements fall into two main categories; the **endogenous retrovirus** (ERV) or long terminal repeat (LTR) retroelements which are similar in structure to retroviruses, and the non-LTR retroelements which are grouped into two distinct families of autonomous **long interspersed repeats** (LINEs), and non-autonomous **short interspersed repeats** (SINEs). Both types of retroelements replicate via an RNA intermediate, which undergoes reverse transcription into DNA and is then inserted semi-randomly into the genome, with the parent element being left in place. Retroelements are often considered to be genomic parasites, since they have no apparent function and a tendency to multiply when they are not controlled by mechanisms such as methylation. Apart from their method of replication, the two types of retroelements have little in common.

The very large numbers of retroelements create problems for anyone intending to work with them. Because members of a given family are similar to each other, it may be difficult to distinguish them; because they are not identical, it may be difficult to design strategies that include every member of a family.

1.1.2 ERV retroelements

ERV elements (including solo LTRs) are present in a very large number of copies in both the human and mouse genomes. There are approximately 450,000 elements in the human genome, which equates to 8.5% of the genome, and 630,000 in the mouse genome, which is 9.9% of the total genome (Venter *et al.* 2001; Waterston *et al.* 2002). There are a range of genome sizes for the ERV families, with the average size being 6-11 kb for full length elements and 500 bp for solo LTRs. Some examples of ERV elements are ERV-9, HERV-K and S71 in humans and IAP, MMTV and VL30 in the mouse.

ERV retroelements are similar to retroviruses – in fact they are homologous. They have two LTRs which flank the element and contain the enhancer, promoter and regulatory signals (refer to Figure 1-1). The LTR consists of three components - a U_3, R and U_5 sequence. The U_3 sequence contains signals to specify and regulate transcription; the R sequence is where transcription is initiated, and it is required for end to end transfer of the growing chain during reverse transcription; the U_5 sequence is thought to contain the signals that regulate translation (Coffin 1992). The size and sequence content of these regions vary between ERV families, as well as within some ERV families, but most elements contain complete LTR regions. There are also a number of solo LTRs scattered throughout the genome which have been formed by recombination of the 5' and 3' LTRs (Bock and Stoye 2000), which are still likely to be capable of active regulation even though they do not possess the protein encoding genes.

Autonomous ERVs contain three genes encoding capsid (*gag*), reverse transcriptase (*pol*) and envelope (*env*) proteins. However most ERVs are missing some or all of the protein coding sequences (Bock and Stoye 2000) which means they are incapable of autonomous retrotransposition. The main difference between ERVs and retroviruses is thought to lie in the envelope coding sequence, which in ERVs allows them to form virus-like particles that however are not infectious – that is they cannot leave the cell, enter another cell, and integrate into its genome (Finnegan 1997).

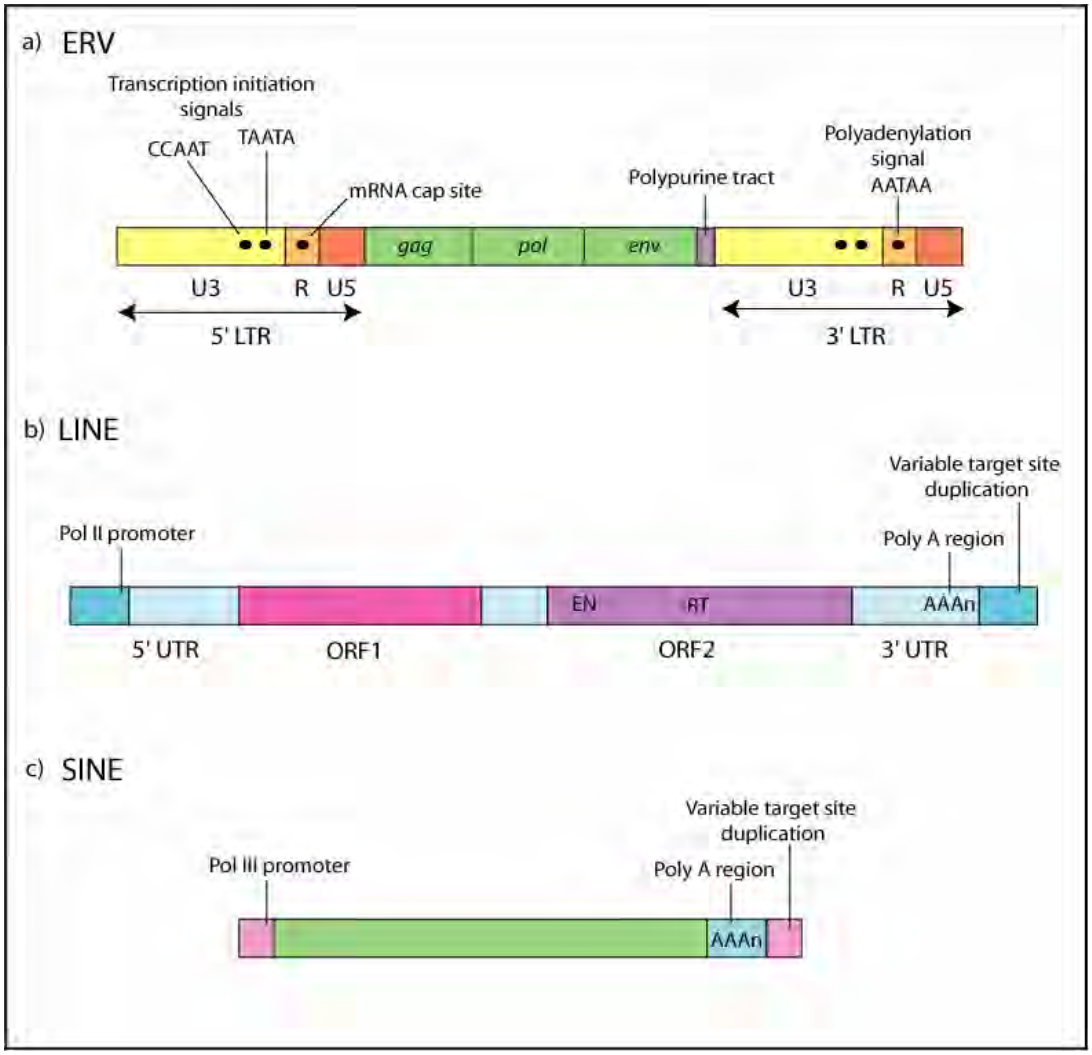

Figure 1-1 Genomic organisation of ERV, LINE and SINE retroelements. **a)** An autonomous ERV element is composed of two LTRs, where the transcriptional regulatory signals are located, and three protein encoding genes ***gag***, ***pol*** and ***env***. There are also numerous solo LTRs throughout the mammalian genome, which consist of a single LTR and they are therefore non-autonomous. **b)** LINE elements consist of a 5' UTR containing an internal polymerase II promoter, two ORFs and a 3' UTR containing a poly A tail. The endonuclease (EN) and reverse transcriptase (RT) positions in ORF2 are shown. LINEs are often flanked by 7-20bp variable target site duplications. **c)** SINE elements consist of a polymerase III promoter and a poly A tail and they are also often flanked by variable target site duplications like the LINEs. Since SINEs do not encode for proteins they are non-autonomous.

Non-autonomous ERV elements and solo LTRs are thought to be capable of transposing with the aid of reverse transcriptases provided by intact ERVs (similar to SINE elements which use LINE components) (Leib-Mosch *et al.* 1993). This dramatically increases the number of possible retrotransposing elements, although it should be noted that ERV elements are thought to be nearly extinct in humans (in other words, there are few or no fully intact ERVs). While the mouse has numerous active members (Venter *et al.* 2001), ERV retrotransposition in the mammalian genome is a rare event. This indicates that mobilisation of intact elements, let alone recruitment of defective elements, is not common.

1.1.3 Non-LTR retroelements

As the name suggests, the non-LTR retroelements do not possess LTRs and have a sequence structure that is completely different to the ERV elements. While LINE and SINE elements are both grouped under this category, they actually have very different sequence structures.

LINE elements have an internal RNA polymerase II promoter, which is located within the 5' untranslated region (UTR), and the promoter sequence is included in the final transcript (Deininger and Batzer 2002). Rodent LINE elements differ from human LINE elements in the 5' UTR, where they have a variable number of tandemly repeated sequences of 205-210 bp called monomers, which are followed by a short non-monomeric region (Ostertag and Kazazian 2001). LINE elements have two open reading frames (ORFs): the first is thought to encode for a protein which acts (like a capsid protein) as a chaperone for the RNA genome (although there is little sequence similarity between LINE subgroups), and the second ORF encodes for a reverse transcriptase and an endonuclease. It is interesting to note that the reverse transcriptase appears to be distantly related to the *pol* gene of the ERVs (Malik and Eickbush 2001),

and it may actually be a precursor to the ERV enzyme. The LINE elements terminate with a poly A tail and they are often flanked by 7-20 bp variable target site duplications (Ostertag and Kazazian 2001). Refer to Figure 1-1.

Like ERV elements, most LINE elements are degenerate and contain deletions, truncations and stop mutations. There are however many full length LINE elements; for example there are 3000-5000 full length L1 sequences in humans and about 3000 full length Tf family LINEs in the mouse (Ostertag and Kazazian 2001).

Of the three retroelement families, the LINE element family makes up the largest percentage of the genome in both the human and mouse, even though they are not present in the largest number of copies. This is likely to be due to their size - an average size for a LINE element is 6 kb. There are approximately 850,000 LINE elements in the human genome, which equates to 21% of the genome, and 660,000 in the mouse genome, equating to 19% of the genome (Venter *et al.* 2001; Waterston *et al.* 2002).

SINE elements do not have any protein coding genes and are therefore non-autonomous. It is thought that SINE elements retrotranspose with the aid of proteins supplied by the LINE elements (Smit 1996; Finnegan 1997; Dewannieux *et al.* 2003). Most SINEs are derived from tRNA, with the notable exceptions of Alu elements, which are derived from the ubiquitous 7SL RNA component of the SRP, and rodent ID elements, thought to be derived from neuronal BC1 RNA (Weiner 2002). Like LINEs, the SINEs have an internal promoter (although it uses RNA polymerase III), they terminate with a poly A tail, and often have similar variable target site duplications (refer to Figure 1-1).

SINE elements are small in size, averaging only 300 bp, making them the smallest of the retroelements, although they are present in the highest number of copies

in both the human and mouse genomes. There are approximately 1,500,000 SINE elements in both the human and mouse genomes and this equates to 13% and 8% of the genome respectively (Venter *et al.* 2001; Waterston *et al.* 2002). Some examples of SINE elements are the Alu elements in humans and the rodent B1 and B2 families.

1.2 Evolution of retroelements

A commonly accepted view of retroelement evolution is that retroelements descended from retroviruses, and that they are degenerate forms of infectious retroviruses which inserted into the genome at some time in the past. During evolution the retroviruses lost their ability to become infectious, which is thought to have occurred via changes to the envelope protein structure, so they were maintained by the host (Coffin 1995).

In this school of thought, the endogenous retroelements are divided into two groups: ancient and modern retroelements. Ancient retroelements inserted into the germline of the ancestral species, and they are therefore not polymorphic in location. They became degenerate over time and are widely dispersed in mammals with members being similar to mammalian type C, B and D viruses. Modern retroelements, on the other hand, are closely related to exogenous retroviruses. They have recently been active (and in some cases are still active), and this is observed by their polymorphic locations within the same species. They are closely related in sequence to each other and they are sporadically distributed, with mice being one of the small number of hosts. Humans are not thought to have any modern retroelements (Coffin 1995).

There is an alternate view of retroelement evolution, however, which directly contradicts the “retrovirus first” view, and proposes that retroviruses evolved from endogenous retroelements. In this view, infectious retroviruses are descended from ERVs that acquired *env* proteins that allow them to leave the cell and enter another cell.

This theory is based in part on studies which have examined the evolution of the reverse transcriptase enzyme (RT) (Temin 1992; Malik and Eickbush 2001). The non-LTR RT in most cases does not encode for a ribonuclease H (RNH) domain; this is thought to be due to the method of target-primed reverse transcription it employs. The non-LTR elements reverse transcribe their RNA template directly onto the chromosomal target site, and it is assumed that the nucleus contains enough RNH activity for this to occur. In contrast ERV elements encode for a RNH domain, and this is thought to reflect the fact that reverse transcription of the ERV RNA template occurs in particles in the cytoplasm where there is likely to be little RNH activity. Malik and Eickbush (2001) used a RNH phylogenetic analysis to show that non-LTR elements were likely to be the precursors to ERV elements, and they suggested that ERV elements were actually fusions between a transposon and a non-LTR element. Retroviruses have RNH domains which appear to have been acquired from non-LTR elements or from a eukaryotic host genome, and the pre-existing domain (as is seen in ERV elements) degenerated to become the tether domain of the RT-RNH complex.

The "retrovirus first" view does not address the evolution of non-LTR elements, only ERV elements, although the similarities between the two families of retroelements may be partially explained by convergent evolution. The "retroelement first" view however does shed light on the evolution of both ERV and non-LTR elements, as well as the evolution of infectious retroviruses, and on the whole seems a more plausible theory of retroelement evolution.

1.3 Epigenetics and retroelements

Mendelian inheritance, in which a given allele of a gene is strongly associated with a particular phenotype and alleles are passed from one generation to the next in undiluted form (Russell 1998), is not always observed in mammals: often variable

phenotypes result from a fixed genotype. The biometrical school proposed that this variability is due to many different genes, each with allelic variability, contributing to the total variability of the observed trait (Risch 2000). Instead of the trait being the result of a single locus, they believe that multiple different loci are likely to be involved and that the trait is therefore complex or "polygenic". These theories do not explain phenotypic variation in genetically identical individuals; in this case variation is usually attributed to environmental factors. If environmental factors are controlled, "intangible variance" is used to describe the phenomenon (Falconer 1989). Recently epigenetics has provided new explanations for non-Mendelian inheritance.

Epigenetics is a term which is used to describe the phenomenon of heritable changes in gene expression which are not due to either changes in DNA sequence or environmental factors (Wolffe and Matzke 1999). This epigenetic variation is thought to be due to modifications such as cytosine methylation, chromatin remodelling, and RNA interference, which act at the transcriptional and posttranscriptional level. Most epigenetic changes are more frequent and readily reversible than genetic changes, and they are more likely to affect somatic than germline cells. Possible examples of epigenetic variation in humans are the differences observed in monozygotic twins. Various studies have shown that environment does not sufficiently account for phenotypic differences in monozygotic twins, and that a component active during embryogenesis is the likely modifying agent (Gartner and Baunack 1981; Gartner 1990; Martin *et al.* 1997). Since genetic variation is negligible in monozygotic twins, epigenetic modifications provide efficient explanations for the observed differences in penetrance and expressivity.

1.3.1 Retroelements as epigenetic modifiers

Transposable elements, or "controlling elements" as they were originally named, were first described by Barbara McClintock in the 1940s after her extensive work with maize (Fincham and Sastry 1974). McClintock (and others working on maize) found that unstable mutations in maize were the result of controlling elements integrating close to or within a gene, thereby altering the activity of the gene. The alteration of surrounding gene expression often manifested itself as phenotypic variation.

The controlling elements McClintock studied are now termed DNA transposons, which move by excision from one locus and reintegration in another. Recent studies in mice have highlighted that retroelements can also act as controlling elements in mammals, with a resultant influence on phenotype. The effects exerted by controlling elements involve ***transcriptional interference***, which we define as **the influence of one transcriptional unit on another unit linked in *cis***. Transcriptional interference is not well understood, but it is likely to involve multiple mechanisms ranging from competition for transcriptional machinery to RNAi. For more detail refer to Section 1.6.1.

A well-studied example of a mammalian retrotransposon acting as a controlling element through transcriptional interference is the agouti viable yellow (A^{vy}) allele in mice. In A^{vy} mice an intracisternal A particle (IAP) ERV retroelement has inserted into pseudoexon 1A of the agouti locus, in an antisense orientation with respect to the gene (refer to Figure 1-2). A cryptic promoter in the 5' LTR constitutively transcribes the agouti gene. Under normal circumstances the agouti gene would be transcribed only during the mid portion of the hair growth cycle, resulting in production of a sub-apical yellow band on a black hair. When the agouti gene is transcribed by the constitutive IAP promoter, it results in mice with completely yellow fur, obesity, type II diabetes

and increased tumours (Duhl *et al.* 1994; Morgan *et al.* 1999). Mosaic activity of the IAP retroelement results in isogenic mice with coat colours that vary from full yellow, to a mixture of yellow and agouti, to full agouti (Duhl *et al.* 1994; Argeson *et al.* 1996; Morgan *et al.* 1999).

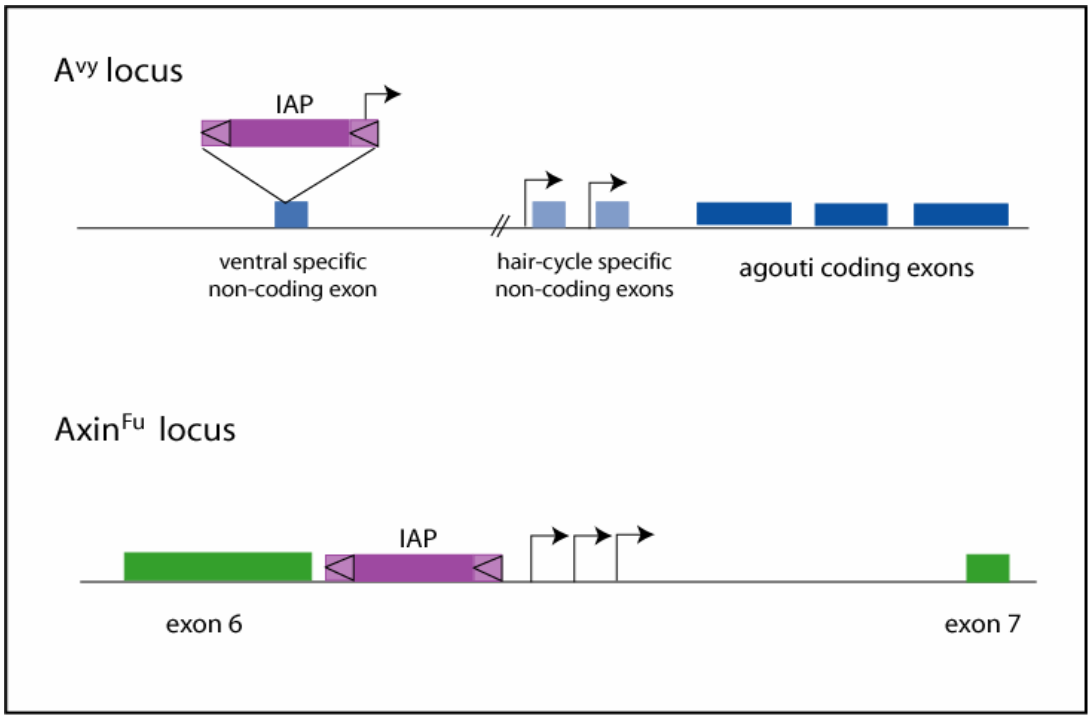

Figure 1-2 Diagrammatic representation of the A^{vy} and $Axin^{Fu}$ loci. The A^{vy} locus has an IAP element inserted into the ventral specific pseudoexon 1A in an antisense orientation. The cryptic promoter drives expression of the agouti gene resulting in constitutive expression of the gene in multiple tissues. The $Axin^{Fu}$ locus has an IAP element inserted into intron 6 of the axin gene and this results in the formation of mutant transcripts initiated in intron 6 itself. Diagrams were adapted from Morgan *et al.* (1999) and Rakyan *et al.* (2003).

Another example of a retroelement acting as a controlling element has been seen in axin fused ($Axin^{Fu}$) mice. Activity of an IAP retroelement, which has inserted in an antisense orientation (relative to the gene) into intron 6 of the axin gene, causes aberrant transcripts which result in a kink in the tail of penetrant mice (Vasicek *et al.* 1997;

Rakyan *et al.* 2003). Refer to Figure 1-2 for a diagram of the locus. The kinky tailed phenotype is caused by axial duplications during embryogenesis, and it is only seen in mice that have an active IAP element. Littermates in which the IAP is epigenetically silenced have a normal tail phenotype. It is interesting to note that the aberrant transcripts do not initiate in the IAP, but rather in intron 6 itself, but the aberrant transcripts are not observed in mice with a silent IAP.

In both A^{vy} and $Axin^{Fu}$, the epigenetic state of the retroelement is partially heritable. In A^{vy} mice it has been found that the maternal phenotype is partially heritable. All dams will produce yellow and mottled pups, but a yellow female will have no agouti pups, mottled females have approximately 9% agouti pups, while agouti females have approximately 20% agouti pups (Morgan *et al.* 1999). The phenotypic contribution of the sire however is not related to offspring phenotype, as all sires produced offspring with the same proportion of yellow (40%), mottled (45%) and agouti (15%) pups.

The $Axin^{Fu}$ phenotype is partially heritable through both the paternal and maternal line. Rakyan *et al*. (2003) found penetrant sires produced 76% of pups with a kinky tail, 24% normal while silent sires produced 60% of pups with a kinky tail, 40% normal. Similarly penetrant dams produced 46% of pups with a kinky tail, 54% normal and silent dams produced 30% of pups with a kinky tail, 70% normal.

While the above are clear examples of retroelements acting as epigenetic modifiers, it is unclear how common this phenomenon is in mammals. In fact it is unclear how many retroelements are capable of transcription, let alone transcriptional interference.

1.4 Incomplete somatic silencing of retroelements

Retroelements are present in very large numbers in the mammalian genome (refer to Table 1-1), and mechanisms to silence them have evolved to minimise deleterious consequences for the host (such as transcriptional interference, and widespread retrotransposition leading to mutations). The process of silencing involves a complex interaction between DNA methylation and chromatin structure (Fuks 2005). There are conflicting reports as to the sequence of events in silencing, but the common view is that histone modification is a prerequisite for DNA methylation, since promoters need to be transcriptionally silent before they are methylated. The whole process is dynamic, and since it is an epigenetic process, silencing may also be reversed.

	Human		Mouse	
	Thousands of copies	Fraction of genome (%)	Thousands of copies	Fraction of genome (%)
LINEs	868	20.42	660	19.20
LINE1	516	16.89	599	18.78
LINE2	315	3.22	53	0.38
LINE3/CR1	37	0.31	8	0.05
SINEs	1,558	13.14	1,498	8.22
ALU/B1	1,090	10.60	564	2.66
B2	-	-	348	2.39
B4/RSINE	-	-	391	2.36
ID	-	-	79	0.25
MIR/MIR3	468	2.54	115	0.57
LTR elements	443	8.29	631	9.87
ERV_classI	112	2.89	34	0.68
ERV_classII	8	0.31	127	3.14
ERV_classIII	83	1.44	37	0.58
MaLRs	240	3.65	388	4.82

Table 1-1 Composition of retroelements in the human and mouse genomes.
Numbers were taken from Venter *et al.* (2001) and Waterston *et al.* (2002).

In retroelement silencing DNA methylation is used to maintain elements in a silent state, and in mammals it has been found that most 5-methylcytosine methylation occurs in repetitive elements (Yoder *et al.* 1997). This methylation however undergoes

changes during embryogenesis, and it has been hypothesised that retroelements may escape remethylation during these changes, resulting in incomplete somatic silencing in a subset of retroelements (Whitelaw and Martin 2001).

1.4.1 What is methylation?

The process of DNA methylation involves the addition of a methyl group to a cytosine residue at the C-5 position as shown in Figure 1-4. 70-80% of all CpG dinucleotides in the human genome have the C residue methylated (Bird 2002), and this methylation has been shown to correlate with gene silencing. Deamination of methyl-cytosines occurs spontaneously, and results in transition mutations of meC to T. The error is often not recognised by cellular machinery, so it is often not repaired (Dean *et al.* 2005). Therefore old methylation sites may be traced by observing alignments of consensus retroelement sequences with the sequences of interest, and recording C-T transitions at CpG sites.

DNA methyltransferases are the enzymes responsible for adding the methyl group to cytosine; there are three family groups designated Dnmt1, Dnmt2 and Dnmt3. The different families appear to have different functions, but as a whole they act to establish the methylated state of DNA by de novo methylation, and then to maintain the methylation once it has been established.

Methylation plays an essential role in X-inactivation, genomic imprinting and silencing of repetitive elements (Bestor 2000). As mentioned most methylation is found in repetitive elements, and because cytosine methylation is heritable it provides a good mechanism for stable silencing of retroelements. As discussed in the following chapters, this silencing is not always complete and some retroelements are capable of becoming active.

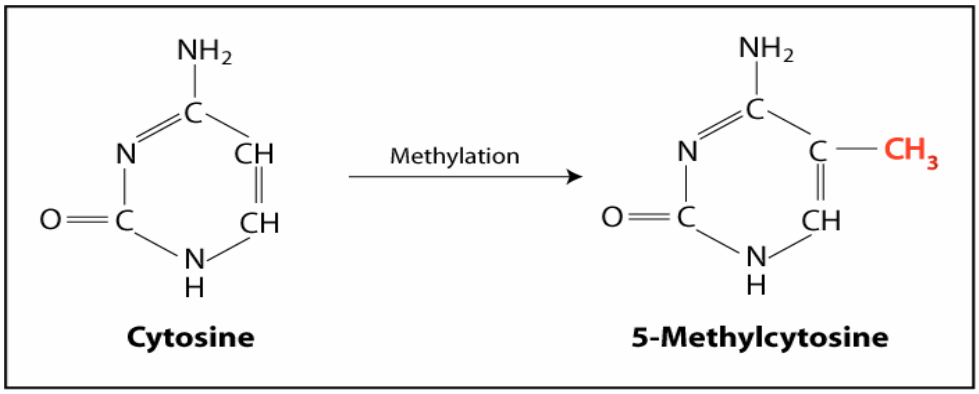

Figure 1-3 Methylation of cytosine to 5-methylcytosine. The methylation step in mammals is carried out by methyltransferases such as Dnmt1, Dnmt2, Dnmt3a and Dnmt3b. The added methyl group is shown in red.

1.4.2 Evidence for methylation of retroelements

Most studies that examine the methylation state of retroelements focus on the L1 element. In L1s there is only one sense promoter to examine, unlike ERV elements which contain two LTRs and so two possible promoters. The L1 element is longer than the SINE elements, which is beneficial for methylation studies as it increases the number of possible CpGs which could be methylated, and it has been extensively characterised. This being said, there are studies which look at other families of retroelements, and the findings suggest that methylation patterns observed in L1 element promoters hold for retroelements in general.

Various methods have been used to examine retroelement methylation, ranging from experimental methods such as bisulfite sequencing (Chalitchagorn *et al.* 2004; Burden *et al.* 2005; Lavie *et al.* 2005) and methylation sensitive restriction digests (Hata and Sakaki 1997; Menendez *et al.* 2004; Suter *et al.* 2004) to computer based methods of examining C-T transitions in sequences of retroelements (DeBerardinis and Kazazian 1999). All have concluded that retroelements are methylated under normal conditions.

In bisulfite sequencing, unmethylated cytosines are converted into uracils by the initial bisulfite treatment. During the PCR step the uracils are read as thymines by the reverse transcriptase. When the PCR products are sequenced, C to T base changes can be observed, and the methylated bases can be determined because they will remain as Cs. Burden and colleagues (2005) used this method to observe CpG methylation in L1 promoters of fetal fibroblasts. An earlier study by Woodcock and colleagues had shown that the L1 promoter was only hemimethylated in human embryonic fibroblasts (Woodcock *et al.* 1997), but Burden *et al*. (2005) showed that hemimethylation of the promoters is uncommon. They actually observed concordant CpG methylation which they also observed in an adult male and adult female fibroblast line as well as in male leukocyte DNA and sperm. In another study which examined the methylation state of L1 elements in normal and carcinoma tissues taken from a variety of organs, it was found that in most tissues examined, the L1 5' UTR region was hypomethylated in the carcinoma tissues compared to the normal counterparts (Chalitchagorn *et al.* 2004). L1 elements are not the only retroelements to be examined by this method: Lavie and colleagues (2005) examined HERV-K 5' LTR methylation states, by combining bisulfite sequencing with reporter construct assays, and showed that methylation levels negatively correlate with transcriptional activity. These are just some of the studies which have successfully used bisulfite conversion to demonstrate the importance of methylation in control of retroelement activity.

Assays based on methylation-sensitive restriction enzymes are not as sensitive as bisulfite sequencing, because while overall methylation patterns can be observed, the proportions of methylated bases in a sample can only be inferred by differences in band intensities; furthermore the enzyme only samples one or a few CpGs in the region of interest. Suter *et al*. (2004) employed methylation sensitive restriction assays to show

L1 promoter sequences were hypomethylated in the colon tumours and normal colonic mucosa of 6/19 colon cancer patients. However this hypomethylation was not observed in normal peripheral blood, lymph node or smooth muscle tissue from these patients, nor was it present in colonic mucosa from healthy individuals. Similarly Menendez and colleagues (2004) found reduced methylation of L1 and HERV-W promoters in ovarian tumours using a similar methylation sensitive restriction assay. They also found expression levels of the two retroelements to be elevated in the tumours when compared to non-malignant ovarian tissue.

Computer-based methods of identifying C-T transitions were used by DeBerardinis and Kazazian (1999) to gain evidence that L1 promoters are three times as likely to have a C-T transition than the L1 body sequence. They interpreted this to indicate that the promoter region is methylated, because otherwise the transition rate in a monomer would not be so high (71%). An interesting suggestion from their conclusion was that these transitions may help the elements escape silencing, but they balanced this with the idea that the transitions may also occur in regions necessary for expression. In any case their finding may explain some cases of elements escaping the methylated state.

Taken as a whole, a variety of methods have been used to show that retroelements are kept silent primarily by methylation, and that when this methylation is removed it leads to increased expression of the elements. Hypomethylation of elements appears to occur only in diseased tissue, or tissue which may appear normal but is in the process of becoming diseased. There is no evidence that the retroelement is causing the diseased state. Rather it appears that the activation of the element is a by-product of cellular deregulation. In any case, for transcriptional interference to occur, the

methylation which is used to silence the retroelements needs to be removed, and there is an opportunity for this to occur in preimplantation embryos.

1.4.3 Methylation changes during embryogenesis

Transcriptional interference requires that retroelements escape silencing and become active in somatic cells. It is has been proposed that this commonly occurs during the resetting of methylation in early embryogenesis, resulting in incomplete resetting of the epigenotype (the genotype with any modifications that alter gene usage) (Whitelaw and Martin 2001).

The patterns of methylation during embryogenesis have been studied in detail in the mouse, where it is possible to obtain samples from the different developmental stages. Therefore the following information pertains to what has been observed in mouse development.

In the maternal and paternal genomes, genomic imprinting of a subset of genes results in methylation differences between the genomes. Approximately 80 genes have been found to be imprinted, although expression profiling suggests the number is higher. The pattern of methylation for imprinted genes is maintained during preimplantation development, but is erased in primordial germ cells (PGCs) (Trasler 2005). Imprinted genes then reacquire methylation at gender specific times during spermatogenesis and oogenesis.

In the mouse, DNA in sperm is relatively undermethylated, when compared to somatic tissue, but it is more methylated than the oocyte (Kafri *et al.* 1992; Monk 1995; Rougier *et al.* 1998). So before fertilisation occurs there are already methylation differences (refer to Figure 1-5). Following fertilisation, demethylation occurs more rapidly in the paternal than the maternal genome, with both genomes reaching a similarly low level following establishment of the blastocyst. De novo methylation occurs at the time of

implantation, and in somatic cells the level of methylation rapidly increases until it peaks at around E5-6.5.

PGCs also undergo demethylation of all single-copy and imprinted genes, primarily from E11.5-12.5. Methylation is reapplied by approximately E18.5 in spermatogenesis and by E15.5 in oogenesis, but the oocyte then undergoes another round of demethylation leaving the methylation in the oocyte lower than that in the sperm.

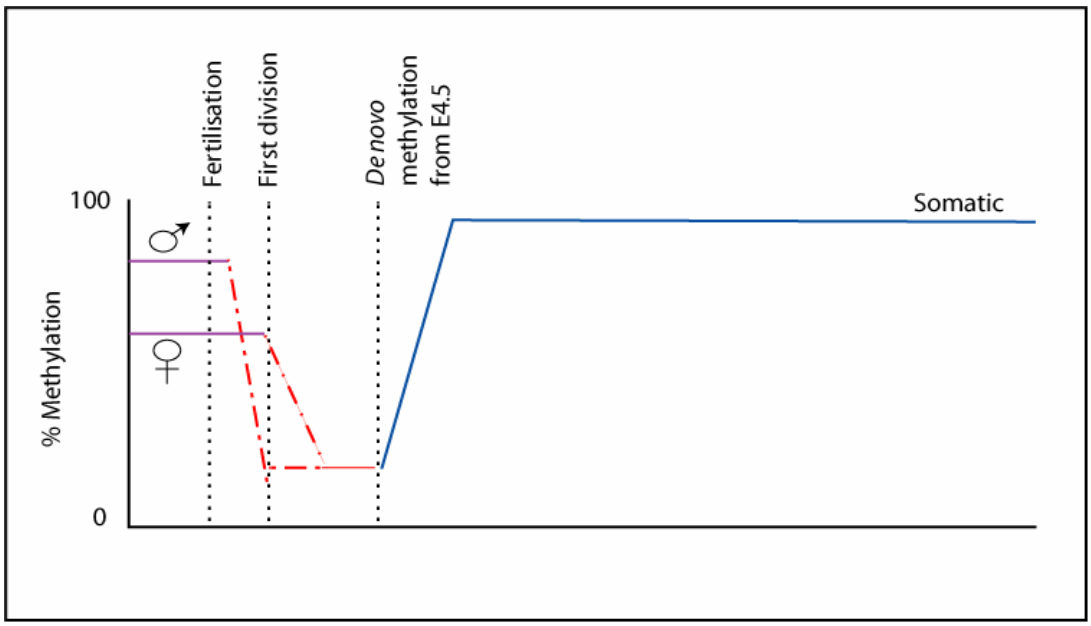

Figure 1-4 Methylation changes over embryogenesis. Purple lines are the methylation states of the paternal and maternal genomes, dashed red lines show active demethylation, the blue line shows remethylation of somatic cells. Extra embryonic and germ line cells (not shown) follow a different path of remethylation.

Therefore during the time period from fertilisation until implantation, retroelements may have an opportunity to escape silencing and become active in somatic cells. There is then an additional opportunity in PGCs, when methylation is reduced and retroelements again may escape the silent state. Since silencing is an

independent and stochastic event for each retroelement, some elements may be missed during the process of remethylation during gastrulation. This opens the doorway for elements to remain active in a mosaic and stochastic manner, resulting in cells within an embryo differing in epigenotypes, depending on the retroelements which have remained active.

1.4.4 Changes in retroelement methylation during embryogenesis

Groups such as Walsh *et al*. (1998), Lane *et al*. (2003), and Kim *et al*. (2004) have studied retroelements, with a focus on the IAP element, to observe the methylation changes which occur in embryogenesis. ETn and L1 elements were also studied by Kim *et al*. (2004) and Lane *et al*. (2003) respectively and they found evidence to suggest that methylation patterns of retroelement families may vary, depending on the family, over embryogenesis.

IAP elements have been found by all groups to be heavily methylated in the sperm and oocyte genomes as well as in the zygote, but by the blastocyst stage a moderate amount of demethylation is observed. Lane *et al*. (2003) observed a fairly high level of methylation of IAP elements in E11.5 primordial germ cells (74%), which was followed by demethylation by E12.5 (40% males and 34% females), with no further demethylation occurring by E13.5 (32% males and 61% females). The demethylation pattern around E11.5 is due to the active demethylation which occurs in germ cells (something which does not occur in somatic cells). Methylation is not re-established in these cells until E16.5-E18.5 (Lane *et al.* 2003). These findings contradict those of Walsh and colleagues who found large scale demethylation in E13.5 embryos (Walsh *et al.* 1998). Lane and colleagues suggested the discrepancy may be due to differences in methods between bisulfite sequencing and methylation sensitive restriction assays (bisulfite sequencing is more sensitive) along with probe differences. All three studies

conclude that IAP elements on the whole are resistant to demethylation during preimplantation development (refer to Figure 1-6), however this does not appear to be the case for all retroelements.

Lane and colleagues (2003) also looked at L1 element promoter methylation and found the changes in methylation were quite different to the IAP elements. L1 element promoters were heavily methylated in sperm (98%), while oocytes showed low levels of methylation (29%). Zygotes and blastocysts also had low levels of methylation (25% and 27% respectively), similar to the oocyte, however *de novo* methylation must occur some time after the blastocyst stage as by E11.5, PGCs were found to have moderately methylated L1 promoters (65%). This methylation dropped in PGCs to a low level by E12.5 (31% males, 34% females) and by E13.5 it had dropped further (13% males, 21% females). Refer to Figure 1-6. This suggests that L1 elements are not resistant to the demethylation process during preimplantation like the IAP elements are.

Kim and colleagues (2004) found similar results for promoter methylation changes of ETn elements. Initially sperm and oocytes were found to have the same amount of moderately high levels of methylation (78%) of ETn promoters, and by the 1 cell stage this had dropped to a moderate level (52%). By the 4 cell stage it had dropped to low levels (23%) and this was maintained into the 8 cell stage (28%). There was an increase in methylation to a moderate level by the morula stage (44%), and this had increased slightly more by the blastocyst stage (49%) (Kim *et al.* 2004). These results also suggest ETn elements are not resistant to demethylation, although the methylation changes they undergo are different to those observed for the L1 promoter (refer to Figure 1-6).

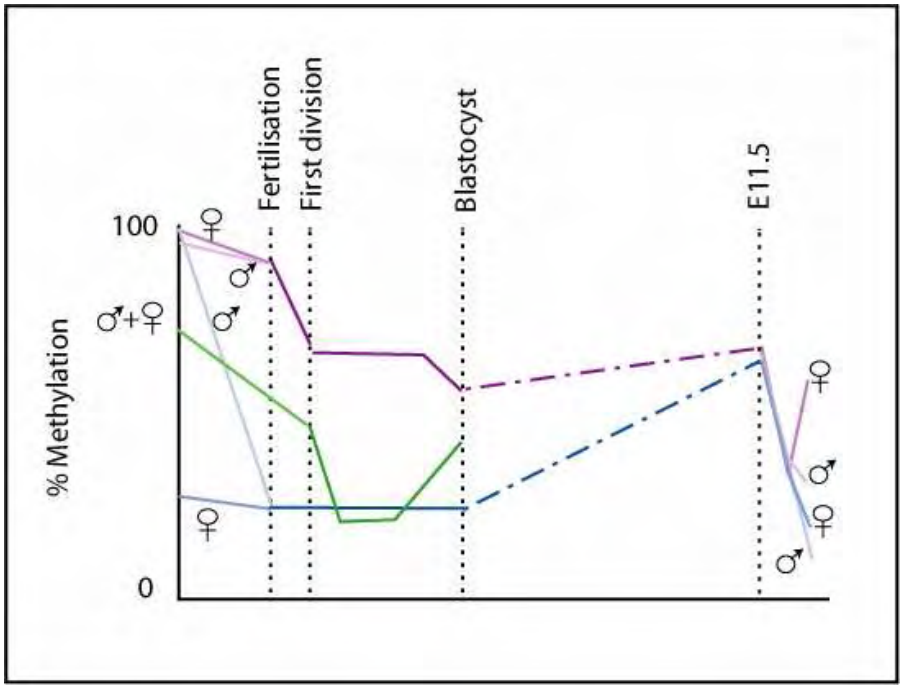

Figure 1-5 Methylation changes in IAP, L1 and ETn elements during embryogenesis. Methylation changes are shown in purple for IAP elements, in blue for L1 elements and in green for ETn elements. The female genome is the darker line in each case except for ETn where both sperm and oocytes had the same methylation levels. The diagram is designed to give an overall pattern and is not to scale, with results being compiled from Lane *et al.* (2003) and Kim *et al.* (2004). There are no results for ETn after the blastocyst period. Dashed lines are where data is not available.

These results show that there are stages when methylation of retroelements is low – ranging from the blastocyst stage for IAP elements to the earlier zygote and 4 cell stages for L1 and ETn elements, and then again in the PGCs from approximately E11.5 for the two families examined. During this time elements would be capable of becoming active. It is also possible that the *de novo* methylation may miss remethylating some of these retroelements, and this would result in a subset of elements which remain active in a stochastic and mosaic pattern in somatic cells. This means there are a subset of retroelements in somatic cells that would be capable of

transcriptional interference. But what evidence is there that this type of scenario is occurring? There have been numerous reports of retroelements becoming active and transcribing and they are looked at in more detail in Section 1.5.

1.4.5 Other silencing mechanisms?

While methylation is thought to be the main method of silencing retroelements in the mammalian genome, there are other silencing mechanisms such as RNAi (Svoboda *et al.* 2004; Soifer *et al.* 2005) and viral editing (Esnault *et al.* 2005) which may also play a role in silencing retroelements.

RNA interference (RNAi) is a silencing mechanism where double stranded RNA (dsRNA) induces sequence specific degradation of homologous mRNAs (Svoboda *et al.* 2004). The process is initiated by dsRNA being cleaved into 21-23 nucleotide silencing RNA (siRNA), by an enzyme known as Dicer, and the siRNA directs sequence recognition and degradation of homologous mRNA. RNAi does not seem to play a role in silencing in adult cells, where it is commonly believed that methylation and chromatin structures help silence unwanted transcription. However there is a potential role for RNAi when these mechanisms are not functioning, i.e. during preimplantation development. Svoboda *et al*. (2004) found evidence of RNAi being employed to silence Merv-L and IAP elements in preimplantation embryos. They inhibited the mDicer pathway and saw upregulation of transcripts from both elements. Another study conducted by Soifer *et al*. (2005), while an *in vitro* assay, also highlighted how L1 elements may be silenced by RNAi. Double stranded L1 RNA was cleaved by DICER (which had been produced by human cells) into short interfering RNA (siRNA). The siRNA was then shown to degrade L1 transcripts, and also suppressed expression of a highly active L1 clone. This suggests that all the mechanisms are in place for RNAi silencing of L1 elements.

Viral editing of retroelements involves mutations of retroelement RNA sequences, via enzymes such as cytidine deaminase, during the reverse transcription step of retroelement replication. The cytidine deaminase enzymes deaminate cytosine residues to uracils, often resulting in degradation of the retroelement DNA, which in turn inhibits retrotransposition. Esnault and colleagues (2005) examined the affect APOBEC3G cytidine deaminase had on retrotransposition of IAP, MusD plus the mouse L1Md element and the human L1H element. APOBEC3G cytidine deaminase induces G-A substitutions in DNA sequences and it has been found to be active in early embryos as well as in germ cells - the times when retroelements are escaping methylation. They found that APOBEC3G impaired retrotransposition of an IAP and a MusD reporter construct, but had no affect on either the human or mouse L1 retrotransposition. When they looked at sequence alignments for endogenous elements they also found that G-A substitutions were observed for MusD and IAP elements but not for L1Md. Esnault and colleagues felt that these results suggest that an APOBEC3G like enzyme is therefore likely to be active *in vivo* and that there may be another mechanism which is used to silence other retroelements such as the L1 elements.

These are two more examples of other components which may be involved in silencing retroelements. The interplay of silencing mechanisms may help explain why only a small subset of the large number of retroelements in the genome, appear to be active at any one time.

1.5 Evidence of retroelement activity *in vivo*

For retroelements to exert transcriptional interference they do not need to be fully intact and capable of transposition, but merely need to have intact promoters which are capable of transcription. While there have been no studies which look at the activity

of a large range of retroelements, there have been reports of retroelement activity, with transcripts being identified in normal as well as diseased tissue.

1.5.1 Retroelement promoters

Eukaryote promoters fall into two main categories – TATA dependent promoters and TATA independent promoters (Pedersen *et al.* 1999; Smale and Kadonaga 2003). TATA dependent promoters, as the name suggests, have a TATA box which is found 25-30 bp upstream of the transcription start site, and this sequence helps position the transcriptional machinery so transcription may occur. However not all genes contain TATA boxes and some have initiator sequences or downstream core promoter element sequences instead which guide binding of transcription factors (Pedersen *et al.* 1999; Smale and Kadonaga 2003).

Retroelements have examples of both types of promoters with ERV elements being TATA dependent (Coffin 1992) and SINE and LINE elements being TATA independent (Smale and Kadonaga 2003). SINE and LINE elements are thought to be TATA independent as they have an internal promoter which is found downstream of the transcription start site, which therefore means they cannot rely on an upstream TATA box for correct positioning of the transcriptional machinery. The RNA polymerase III promoter region, Box A and B (see Figure 1-7), of SINE elements is thought to help position the transcription factors in SINE elements but LINEs do not have these sequences. It has been speculated for LINEs that it is likely that they use a initiator element (Inr) (Ostertag and Kazazian 2001) or a downstream promoter element (DPE) (Smale and Kadonaga 2003), although the precise sequence they utilise is unknown.

```
                    A Box
Alu con   1 GGCCGGGCGCGGTGGCTCACGCCTGAATCCCAGCACTTTGGGAGGCCGAGGCGGGAGGATCACTTG
AluJo     1 GGCCGGGCGCGGTGGCTCACGCCTGAATCCCAGCACTTTGGGAGGCCGAGGCGGGAGGATTGCTTG
AluSc     1 GGCCGGGCGCGGTGGCTCACGCCTGAATCCCAGCACTTTGGGAGGCCGAGGCGGGCGGATCAC--G
AluSp     1 GGCCGGGCGCGGTGGCTCACGCCTGAATCCCAGCACTTTGGGAGGCCGAGGCGGGCGGATCACCTG
AluSx     1 GGCCGGGCGCGGTGGCTCACGCCTGAATCCCAGCACTTTGGGAGGCCGAGGCGGGCGGATCACCTG
AluY      1 GGCCGGGCGCGGTGGCTCACGCCTGAATCCCAGCACTTTGGGAGGCCGAGGCGGGCGGATCAC--G
AluYb8    1 GGCCGGGCGCGGTGGCTCACGCCTGAATCCCAGCACTTTGGGAGGCCGAGGCGGGTGGATCAT--G
AluYd2    1 GGCCGGGCGCGGTGGCTCACGCCTGAATCCCAGCACTTTGGGAGGCCGAGGCGGGCGGATCAC--G

                         B Box
Alu con  68 AGCCCAGGAGTTCGAGACCAGCCTGGGCAACATAGTGAAACCC 110
AluJo    68 AGCCCAGGAGTTCGAGACCAGCCTGGGCAACATAGCGAGACCC 110
AluSc    66 AGGTCAAGAGATCGAGACCATCCTGGCCAACATGGTGAAACCC 110
AluSp    68 AGGTCGGGAGTTCGAGACCAGCCTGACCAACATGGAGAAACCC 110
AluSx    68 AGGTCAGGAGTTCGAGACCAGCCTGGCCAACATGGTGAAACCC 110
AluY     66 AGGTCAGGAGATCGAGACCATCCTGGCTAACACGGTGAAACCC 110
AluYb8   66 AGGTCAGGAGATCGAGACCATCCTGGCTAACAAGGTGAAACCC 110
AluYd2   66 AGGTCAGGAGATCGAGACCA------------CGGTGAAACCC 96
```

Figure 1-6 Alignment of Alu promoter regions. Alu promoter regions for subtypes AluJo, AluSc, AluSp, AluSx, AluY, AluYb8 and AluYd2 are shown aligned with the Alu consensus sequence (Alu con). Base pair differences are highlighted in red and dashes indicate deletions. The RNA Polymerase III A and B Box elements are shown.

Another function of eukaryote promoters is to direct cell/tissue specific transcription and this is determined by factors such as the presence of different binding sites for different transcription factors (see Figure 1-8). Retroelement promoters have also been shown to have this ability. For example the SOX family of transcription factors are thought to drive specific transcription of L1 retroelements in the germ line (Ostertag and Kazazian 2001) as well as in neuronal precursor cells (Muotri *et al.* 2005). There are in fact numerous reports of transcription factor binding sites which have been reported to be close to or within the promoter regions of retroelements. For instance NFY, M2F1 and GATA-2 sites have been found within ERV-9 LTRs (Yu *et al.* 2005), RUNX3, SOX and YY1 sites have been identified close to LINE promoter regions (Tchenio *et al.* 2000; Yang *et al.* 2003; Athanikar *et al.* 2004) and Sp1 and YY1 sites

within Alu elements (Oei *et al.* 2004). Note that the small size of Alu elements means the binding sites may fall anywhere within the element and they can still affect the promoter. There are numerous reports of retroelements being expressed in a cell or tissue specific manner (Brulet *et al.* 1983; Britten 1996; Peaston *et al.* 2004), however definitive studies have yet to be done to conclusively show these binding sites are driving the selective expression.

Another feature of retroelement promoter regions, which has been shown to drive coordinated expression of retroelements and which may contribute to cell/tissue specific expression of retroelements, is the presence of sequences responsive to hormones, calcium, and other effectors (Leiter *et al.* 1986; Stavenhagen and Robins 1988; Schiff *et al.* 1991; McHaffie and Ralston 1995; Vansant and Reynolds 1995; Morales *et al.* 2002). The effectors have been studied more thoroughly than the transcription binding sites and they have clearly been shown to influence expression of the retroelements.

Retroelement promoters are therefore quite complex and it seems likely that it is a combination of the methylation state of the region, regulatory elements such as enhancers and silencers within the promoters, the presence of effector sequences, as well as the binding of transcription factors that leads to the specific expression of the retroelements. At present most of the information regarding retroelement promoters is merely speculation as there is a relative paucity of information on retroelement promoters, excluding ERV elements where information has been gained from studies of retroviral promoters, due to the fact that most retroelements are maintained in a silent state and are not active. Whatever the factors influencing expression of retroelements, large numbers have been reported to be active and this is discussed further in the following sections.

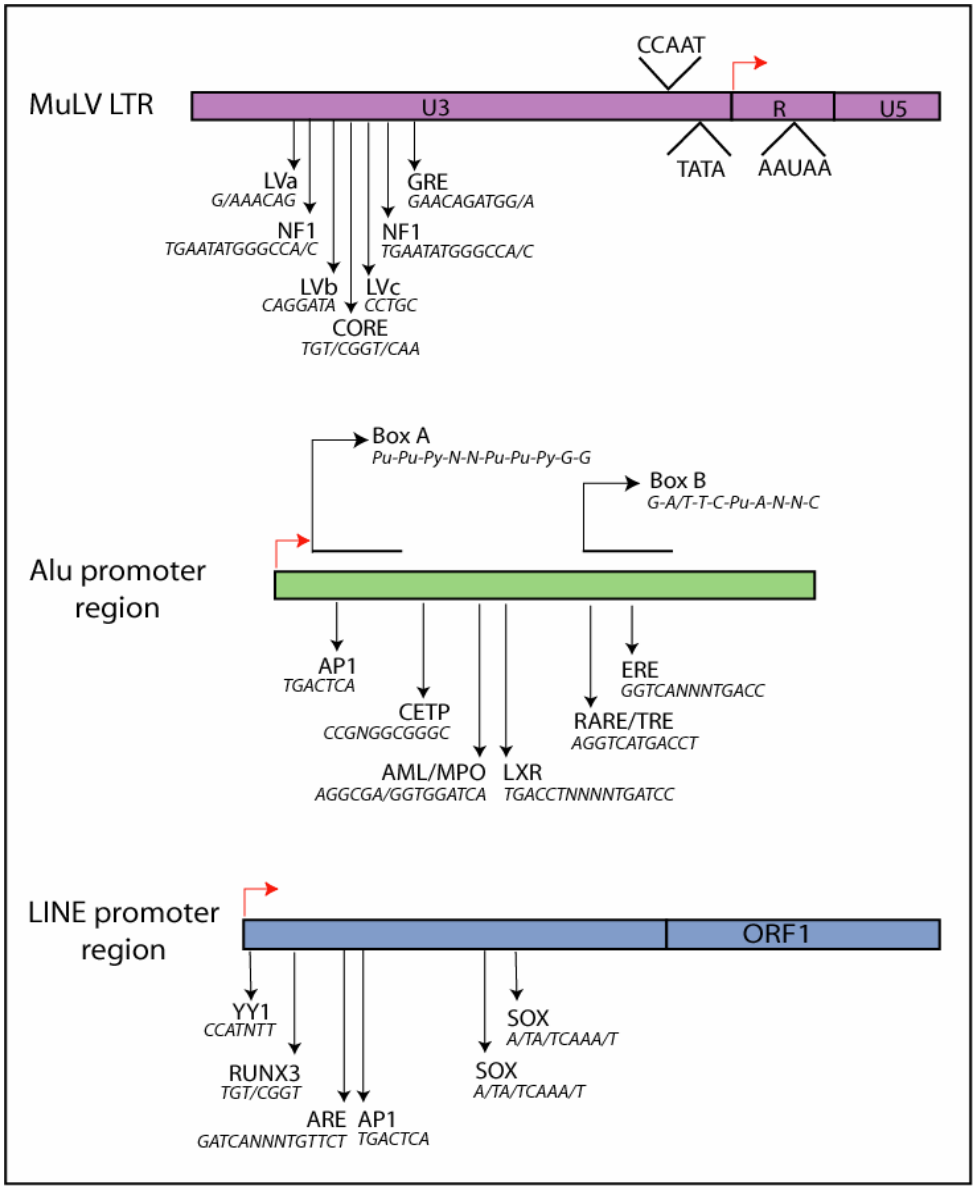

Figure 1-7 Retroelement promoter regions showing transcription factor binding sites and binding sites for other effector molecules. Examples of promoter regions from the three retroelement families are shown above with some of the known transcription factor and effector molecule binding sites. MuLV is a representative of the ERV family - the TATA box, CAAT box and 3' processing sequence (AAUAAA) are shown in addition to the other binding sites. The Alu element is a representative of the SINE family and the location of the RNA Polymerase III promoter regions Box A and B are shown in addition to the other binding sites. The red arrows denote the start site of transcription. There are likely to be more binding sites and not all binding sites shown are necessarily present in all the members of that family of retroelements – the Figure is only a representation of common sites. Adapted from Coffin 1992 and Grover 2005.

1.5.2 Expression of retroelements in embryogenesis

As mentioned in Section 1.4 it is thought that demethylation of the genome occurs in early embryogenesis, which paves the way for activation of retroelements. Therefore it is unsurprising that there are a number of reports of retroelement activity from this period in mammalian development.

Reports on retroelement expression (defined in this case as being transcriptional expression since we are interested in the ability of retroelements to exert transcriptional interference) during embryogenesis come from mouse studies where it is easy to obtain material. Transcription of retroelements has been shown to occur as early as in 2 cell embryos (Piko *et al.* 1984; Poznanski and Calarco 1991) and to extend into the late blastocyst stage (Piko *et al.* 1984; Poznanski and Calarco 1991; Packer *et al.* 1993). Transcription is also detected in embryos as well as in placental tissue (Norton and Hogan 1988) a finding which is supported in human samples as well (Kjellman *et al.* 1999).

Both ERV and non-LTR retroelements appear to be expressed during embryogenesis, however only some of the larger families have been examined and there is a paucity of data on other retroelements. IAP elements were found to be expressed in 2 cell embryos, 8 cell embryos and blastocysts by Piko *et al*. (1984), Poznanski and Calarco (1991) and Svoboda *et al*. (2004) and the number of transcripts appeared to increase throughout preimplantation development. Interestingly Svoboda and colleagues found this pattern was not followed for Merv-L where transcript levels peaked at the 2 cell stage and decreased so they had basically ceased being detected by the blastocyst stage. Packer and colleagues (1993) found L1 transcripts to be abundantly expressed in blastocysts while VL30 was found to be expressed at quite a high level in 10.5-14.5 day embryos. This supports the large scale studies in Section

1.4.4 and suggests that while retroelement expression is common for different families, their expression patterns may be different. This is likely to be due to local chromatin conditions as well as the ability of retroelement expression to be regulated (refer to Section 1.8.2 for more detail).

Less information is available for humans, although there is abundant literature which examines expression of the retroelements HERV-R, HERV-W, and HERV-FRD which are thought to be involved in villous cytotrophoblast differentiation (de Parseval *et al.* 2003; Frendo *et al.* 2003; Rote *et al.* 2004). A quick search of the dbEST will also confirm the expression of other families of retroelements in placental tissue, which suggests that retroelement expression is likely to follow similar patterns to that seen in mouse embryogenesis.

1.5.3 Expression of retroelements in normal healthy adult tissue

Retroelement expression is as common in normal healthy adult tissue as it is in embryogenesis, which seems to contradict the common view that retroelements are maintained in a silent state once embryogenesis is complete.

Expression of both ERV and non-LTR retroelements has been reported in many types of tissues in the mouse. Schiff and colleagues (1991) found expression of VL30 elements in steroid-producing cells in the ovary, testes, adrenal gland and placenta of C57/BL mice. Interestingly the expression appeared to be increased in response to trophic hormones, which suggests that VL30 expression can be regulated. Branciforte and Martin (1994) examined L1 expression in the testis of adult C57BL/6 mice with the focus of showing how germ line mutations via transposition events may occur. They found L1 transcripts in both germ and somatic cells with expression varying with age and cell type (Branciforte and Martin 1994). Dupressoir and colleagues (1995) found IAP transcripts in various organs of 4 week and 22 month old mice. The patterns of

expression were the same for both age groups for the brain, heart, intestine, kidney, lung and spleen but there was an age related increase in expression in the liver. The IAP expressed in the liver was later identified and found to be the result of a single genomic locus becoming activated during aging (Puech *et al.* 1997). A study conducted by Gaubatz and colleagues (1991) at an earlier time found contradictory results to Dupressoir's group. They found IAP expression to be highest in the latter stages of gestation and this remained high in neonatal and 2 month old mice. Expression in heart and liver samples decreased over the 2-32 month period, but remained fairly constant in the brain and kidney (Gaubatz *et al.* 1991). The differences observed in IAP expression between the two groups may be explained by the different mouse strains having a different subset of retroelements active, leading to differences in expression over time. It is more likely however that the differences in expression patterns are due to the two groups using different probes. The probes may have been complimentary to different IAP sequences which may result in the different findings.

These are a few examples of active retroelements in normal mouse tissues. It seems likely that most retroelement families, if not all, are capable of transcription. This supposition is based on the reasoning that if one ERV or LINE family is capable of activity, then the similarity between that family and other ERV or non-LTR members suggests that similar mechanisms would allow for other families to become active.

Reports of retroelement expression in normal healthy adult mouse tissue are scarce, but they are even less common in human studies where normal tissue has only been studied so that it can be included as a control in diseased tissue studies. There are human studies which look at retroelement expression in cell culture RNA; however these are transformed cells that are likely to have expression abnormalities, and so I have not classified them as being normal and healthy.

1.5.4 Age-related retroelement expression

The process of aging is associated with cellular dysfunction and general cell degeneration (Vojta and Barrett 1995). Decreased methylation is one of the parameters which is correlated with aging, and this means that retroelements are presented with another opportunity to escape silencing. There are not many studies which specifically examine retroelement expression in aging tissue, but the few which have looked find retroelement expression to be increased in aged samples.

A study which examined MuLV expression in C57BL/6J mice found that expression of the retroelements, measured by sequence complexity assays, was elevated in aged mouse brain and liver tissue (Florine *et al.* 1980). Using a similar method Dean and colleagues found that MMTV RNA sequence complexity was increased in the liver but not the brain of aged C57BL/6J mice (Dean *et al.* 1985). Taken together these studies suggest that age-related release of retroelements from silencing is not total, and that it may even be specific to certain tissues and retroelement families. In fact a thorough study conducted by Barbot and colleagues (2002) found that the tissue-specific expression of an IAP element in the mouse liver (Dupressoir *et al.* 1995) was due to progressive demethylation associated with repetitive induction. The IAP is located within an intron of the circadian gene, m. nocturnin, and the gene is expressed in a circadian pattern in the liver. The repetitive activation of the gene over time results in the activation of the IAP element in the liver (Barbot *et al.* 2002). It is highly possible that other examples of retroelements escaping silencing during aging may be due to similar positional affects, which could explain the observed tissue specific activation versus global activation of retroelements. It can also be speculated that it seems likely that the retroelements themselves do not bring about aging but rather are expressed due to the cell deregulation brought on by aging.

1.6 Retroelements and transcriptional interference

The majority of studies relating to retroelements focus on the ability of these elements to transpose and therefore cause insertional mutations, exon shuffling, gene breaking and the production of pseudogenes (Moran *et al.* 1999; Esnault *et al.* 2000; Ostertag and Kazazian 2001; Wheelan *et al.* 2005). These studies focus on transposition of retroelements, as it is thought that disease phenotypes are likely to occur only through retroelements transposing and disrupting a gene locus. Few studies look at transcription of retroelements and their ability to exert transcriptional interference on nearby genes, because this property of retroelements is not well understood and it is hard to investigate. Nonetheless it has been shown via studies conducted on the A^{vy} and $Axin^{Fu}$ mice that retroelement-mediated transcriptional interference does occur in mammals, and considering the large numbers of retroelements in the mammalian genome, it is highly likely that these two cases are not isolated.

1.6.1 Transcriptional interference

As described earlier, **we define transcriptional interference as being the influence of one transcriptional unit on another unit linked in *cis***. The interference may occur over great distances (the IAP element is located 100 kb upstream of the agouti gene in the A^{vy} locus) and since transcriptional activity/silence is an epigenetic phenomenon, interference may be reversible and mosaic. The mechanism of transcriptional interference is not well understood but it is likely to include multiple distinct mechanisms such as competition for transcriptional machinery, interference with spatial relationships of enhancers and promoters, antisense transcription which may lead to RNA silencing, readthrough transcription and insulation. Refer to Figure 1-8.

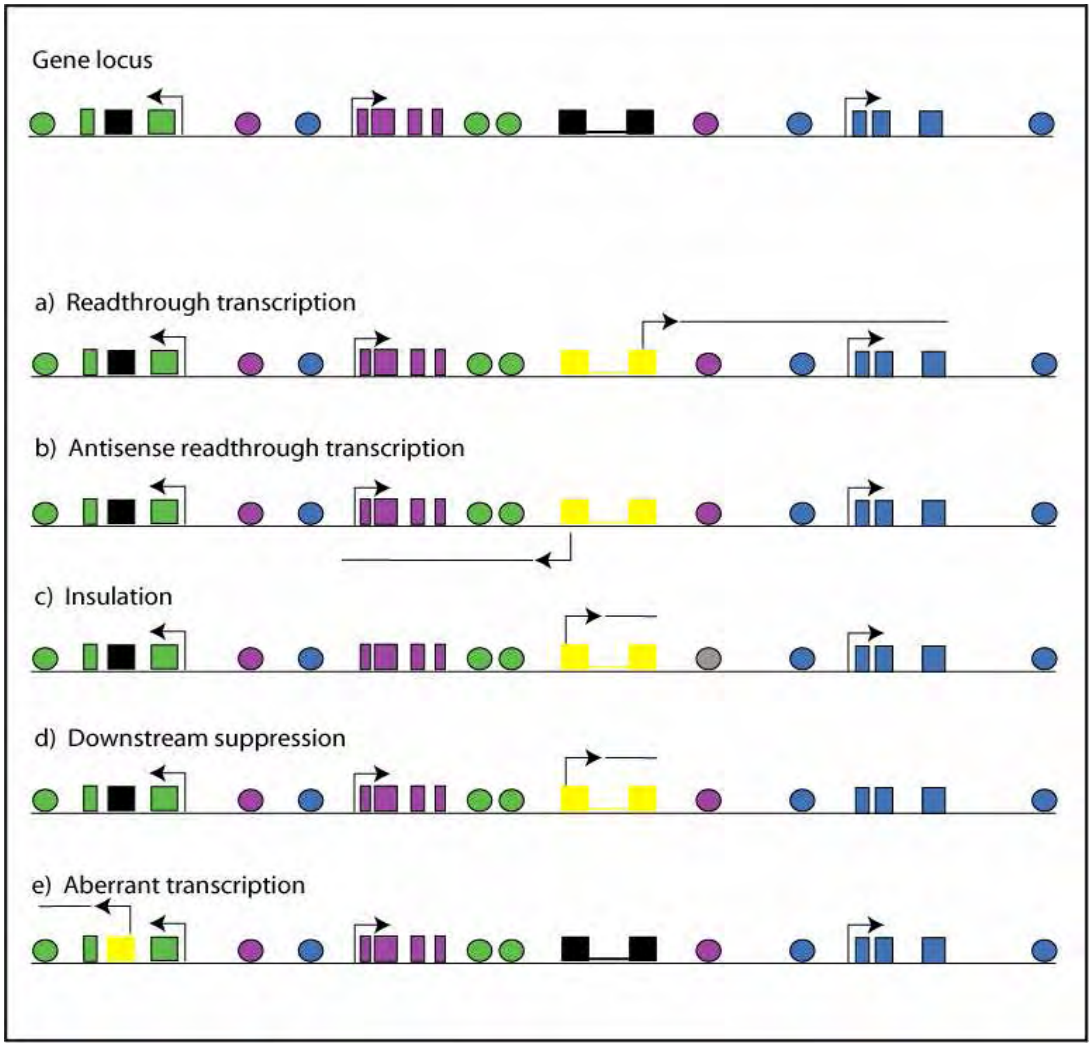
Gene locus
a) Readthrough transcription
b) Antisense readthrough transcription
c) Insulation
d) Downstream suppression
e) Aberrant transcription

Figure 1-8 Examples of transcriptional interference. The examples shown use an ERV element causing transcriptional interference but a non-LTR retroelement could be substituted. A full length ERV element is shown as black boxes in the silent state and yellow in the active state. A solo LTR is shown as a single box. Circles represent enhancers and coloured boxes are exons. Arrows indicate active promoters. **a)** Readthrough transcription is where the retroelement becomes active and transcribes through the element's stop signal into a downstream gene. This may cause the gene to be upregulated or to be expressed in an abnormal pattern. The A^{vy} locus is an example of this kind of interference. **b)** Antisense readthrough transcription is similar to the above example except the retroelement transcribes the antisense strand of the gene. It is thought that this type of transcription could lead to gene silencing via RNAi methods. **c)** Insulation is where an enhancer of a gene is blocked from interacting with the gene's promoter by the insertion of a specialised DNA sequence between the enhancer and promoter. Insulation can have varying results however it most commonly keeps expression of a gene at levels which are not influenced by enhancer action. The Drosophila gypsy retroelement is an example of an element which acts as an insulator. **d)** Downstream suppression is where the activity of a retroelement causes a gene to be silenced downstream. It is likely that competition for transcriptional machinery mediates this type of interference. **e)** Aberrant transcription is where a retroelement which has inserted into a gene – usually into an intron, becomes active and transcribes into nearby exons resulting in mutant transcripts. This is similar to the case of $Axin^{Fu}$ except in that case the retroelement does not provide the promoter. There are likely to be other examples of transcriptional interference – these are just the most common examples.

As mentioned in Sections 1.1.2 and 1.1.3, many retroelements possess promoters which are capable of transcribing – and since an active transcription unit is the main requirement for transcriptional interference, any retroelement with an intact promoter has the potential to exert interference on nearby genes. In plants this is not a new concept, as Barbara McClintock and others first observed this type of interference whilst examining transposons in maize 60 years ago (Fincham and Sastry 1974). Groups studying Drosophila have also widely accepted the fact that transposons and retroelements may perturb expression of surrounding genes (Parkhurst and Corces 1986; Williams *et al.* 1988; Corbin and Maniatis 1989). However in mammals very few examples have been described and the idea of retroelements interfering with transcription of genes has not been embraced.

1.6.2 Evidence of retroelements interfering with gene expression

The A^{vy} and $Axin^{Fu}$ mice are two well-characterised examples of transcriptional interference, both of which were found because of their striking phenotypic effects, which differ between genetically identical mice (see Section 1.3.1). Unfortunately phenotypic changes brought about by transcriptional interference are not likely to be obvious in every case, which may complicate the identification of additional examples.

Retroelements that form chimeric sequences with nearby cellular gene sequences, via either readthrough transcription or splicing mechanisms, are obvious candidates for transcriptional interference. The lab of Mart Speek has looked extensively at the human L1 element; they have characterised an antisense promoter which is located in the 5' UTR and is capable of becoming active (Speek 2001; Nigumann *et al.* 2002). By examining an NTera2D1 library (an embryonal carcinoma cell library) they found not only that the L1 antisense promoter was capable of producing transcripts, but that in some cases the transcripts were spliced to cellular

genes (Speek 2001). This led to an expressed sequence tag database (dbEST) search where they uncovered a large number of chimeric transcripts in both normal and tumour libraries (Nigumann *et al.* 2002). It is interesting to note that the L1 elements were degenerate examples with incomplete ORFs – elements which would be overlooked by other studies which assume that only full length elements should be of interest as transposition is the only important mechanism carried out by retroelements.

Peaston *et al*. (2004) found a similar type of interference in mice when they studied retroelement expression in oocytes and preimplantation embryos. They found that while the mouse transcript (MT) retroelement was the most abundantly expressed element, there were a number of elements acting as alternative promoters and first exons for a subset of cellular genes. In other words the sequence 5' to where the retroelement was inserted was missing from the transcript, and in the cases where the element was located upstream of the gene, one or more conventional exons were omitted allowing for the retroelement promoter to drive expression (Peaston *et al.* 2004). This is an obvious example of retroelements interfering with expression of nearby genes.

Chimeric transcripts are not the only type of transcriptional interference which can easily be observed. As part of a study investigating RNAi in preimplantation mouse embryos, antisense transcription of retroelements has been shown to occur by Svoboda *et al.* (2004). They found that Merv-L and IAP elements were expressed in a sense and antisense direction in 2 cell and 8 cell embryos and when the RNAi pathway was inhibited by silencing mDicer, a 50% increase in expression of Merv-L and IAP elements was observed in 8 cell embryos. RNAi is a form of transcriptional interference which has received a lot of attention over the past few years, and the above study suggests it may play a role in retroelement silencing in the early stages of

mammalian development before other protective mechanisms are established (refer to Section 1.4.5).

These are a few examples which illustrate that transcriptional interference by retroelements may actually be a common phenomenon in mammals. It is just a matter of finding a method to uncover the interference. It can be difficult to search for elements acting, for example, as insulators or by competing for transcriptional machinery, when no abnormal transcription or phenotypic difference is observed. Until this problem is overcome, the link between transcriptional interference and disease is confined to examples like the A^{vy} and $Axin^{Fu}$ mice.

1.7 Retroelements and disease

There are numerous reports dealing with retroelement expression and disease. Many reports are from the perspective of retroelements causing transpositional mutations, but others just report on the observation that retroelement transcripts are associated with disease states. The common idea is that the retroelement activity may have helped bring about the disease, but in fact in my opinion it is more likely to be the reverse: that deregulation brought about by the disease has caused retroelement activity.

A large number of human diseases have been linked to retroelement activity ranging from autoimmune diseases such as diabetes (Badenhoop *et al.* 1996; Marguerat *et al.* 2004) and rheumatoid arthritis (Nakagawa *et al.* 1997; Ali *et al.* 2003) to heart disease (Sirokman *et al.* 1997; Bing *et al.* 1998), schizophrenia (Yolken *et al.* 2000; Karlsson *et al.* 2001) and cancer (Menendez *et al.* 2004; Buscher *et al.* 2005). Apart from studies which focus on diseases caused by retroelement insertions, there has been no real evidence to suggest retroelements have been the causative agents in disease progression. In fact attempts to isolate infectious retroviruses from diseased patients

have failed (Murdoch *et al.* 1990; Krieg *et al.* 1992) and have not been attempted in most studies.

The significance of retrotransposon expression may lie in the ability of these elements to perturb expression of surrounding genes, but studies focused on the impact of retroelements on disease have not examined the diseases from this viewpoint.

1.7.1 Retroelements and type 1 diabetes

HERV-K retroelements are of particular interest to groups studying autoimmune disease as they are thought to encode for a T-cell superantigen (SAg). The superantigen acts by stimulating specific T-cells, which express a SAg receptor, without the presence of antigen presenting cells. This leads to an overstimulation of the immune system with massive cytokine secretion which is disproportionate to the "pathogen" load. The end result is that a number of cells produce excessive inflammation - as is seen in autoimmune diseases. There has yet to be a study which provides evidence that HERV-K proteins elicit such a response so the link is tenuous at best.

Type 1 diabetes is an autoimmune disease which is caused by the host immune system attacking the insulin producing pancreatic β cells. HERV-K elements are found in large numbers in the HLA region and are present in haplotypes associated with diabetes. Marguerat and colleagues (2004) used a large family based association study and found evidence to suggest that the HERV-K18/CD45 locus is associated with type 1 diabetes progression. Earlier Badenhoop *et al*. (1996) also found evidence that a HERV-K LTR in the DQ region segregated with haplotypes which show a predisposition for insulin dependent diabetes mellitus. The same group later found an additional HERV-K LTR (LTR13 subfamily) to be linked with distinct HLA DQB1 haplotypes which confer susceptibility to diabetes (Bieda *et al.* 2002). However none of the studies have experimental evidence to show that the HERV-K elements contribute to

the progression of diabetes, or for that matter that they are even active, so no firm link has yet been established between HERV-K expression and diabetes.

1.7.2 Retroelements and rheumatoid arthritis

Rheumatoid arthritis is another autoimmune disease where retroelements have been found to segregate with haplotypes which show a predisposition for disease progression. The disease is characterised by chronic inflammation of the synovium which results in joint damage and subsequent disability. Seidl *et al*. (1999) examined a HERV-K LTR at the HLA DQB1 locus and found that it was overrepresented in patients versus controls, suggesting that it enhances susceptibility to rheumatoid arthritis (Seidl *et al.* 1999). Nakagawa and colleagues (1997), on the other hand, used an experimental approach to look at retroelement contribution to rheumatoid arthritis progression. The study looked at rheumatoid arthritis, osteoarthritis and normal synovial tissue and found the frequency of sequences from the retroelement families was the same. However when they looked at expression in synovial fluid cells from rheumatoid arthritis sufferers, they found that while some retroelement families were expressed at a lower rate, the ERV-9 related, HERV-K related and HERV-L related families were expressed at a higher frequency than seen in synovial tissue from the same patients. Ali *et al*. (2003) similarly found higher expression of three homologous L1 element transcripts in rheumatoid arthritis synovium compared to the synovium from reactive arthritis patients. Conclusions from the latter two studies were simply that retroelements were being expressed and that further work was necessary to determine if the expression of the retroelements leads to the diseased state. Therefore once again it has not been shown that the retroelement causes the disease.

1.7.3 Retroelements and heart disease

Spontaneously hypertensive rats have been used for many years as models for genetic hypertensive heart disease, therefore it has been of interest to some groups to see if retroelements may be expressed in differing levels in these rats compared to normal controls. Sirokman and colleagues (1997) used a differential display to look at transcript expression in spontaneously hypertensive rats (SHR) and found expression of an ERV element to be greater than an order of magnitude increased in the SHR rats, compared to age matched normal Wistar-Kyoto rats. The retroelement expression level was higher still in SHR with heart failure. The ERV expression was localised to myocardial cells suggesting a link between hypertensive heart disease and expression of the ERV. A later study by the same group (Bing *et al.* 1998) extended the initial study to show that the expression is not secondary to pressure overload, as Wistar-Kyoto rats which developed hypertrophy due to aortic banding did not express the retroelement. Crossing of the SHR and Wistar-Kyoto rats resulted in progeny which had intermediate hypertrophy and expressed the ERV element, suggesting that the expression of the ERV segregates with the SHR genotype. No direct evidence was shown for the retroelement being the disease agent.

1.7.4 Retroelements and schizophrenia

Schizophrenia is a complex brain disease of uncertain aetiology. There is a definite genetic component to the disease, but there have been no genes identified which have a really strong association with disease acquisition (Craddock *et al.* 2006). In an attempt to understand the disease some groups have looked at retroelement expression in cerebrospinal fluid from patients with schizophrenia, as well as frontal cortex samples from post mortem brains. Both Yolken *et al*. (1999) and Karlsson *et al*. (2001) found expression of HERV-W to be higher in samples from schizophrenic patients

versus healthy controls, while ERV-FRD expression was also moderately higher in schizophrenic samples. Interestingly Yolken and colleagues found bipolar patients had upregulated levels of HERV-K10 compared to schizophrenics and normals, suggesting different retroelement expression profiles may apply to different neurological disorders. Only transcripts of retroelements were identified in both studies which means no direct link to the retroelement causing the disease was shown.

1.7.5 Retroelements and cancer

Retroelement expression is commonly linked to cancer and there are numerous studies which look at this association. A summary of some recent studies may be seen in Table 1-2

As with the previous studies relating to retroelements and disease, the retroelements in the above studies have not been directly linked to disease acquisition. There is upregulation of retroelement transcripts, due to the widespread hypomethylation (refer to Section 1.4.2) of most retroelements in cancer cells, but otherwise no direct evidence has been shown to suggest the retroelements examined were causing the disease state. There are two exceptions. The first exception is where Alu elements have been shown to have a direct involvement in cancer progression through genome rearrangements (Deininger and Batzer 1999). Due to the Alu elements being present in such large numbers, genomic rearrangements through unequal recombination and Alu mediated defective splicing can result in disease. L1 elements are also capable of unequal recombination (Burwinkel and Kilimann 1998) but it is a rare event and therefore is not a major cause of disease like Alu unequal crossing over.

Retroelement	Cancer tissue/cell line	Paper Reference
HERV-E	Prostate adenocarcinoma	(Wang-Johanning *et al.* 2003)
HERV-F	Various cancer cell lines	(Yi and Kim 2004)
HERV-H	Various cancer cell lines	(Yi *et al.* 2006)
	Colorectal adenoma	(Wentzensen *et al.* 2004)
HERV-K	Melanomas and melanoma cell lines	(Buscher *et al.* 2005)
	Breast cancer cell lines	(Ejthadi *et al.* 2005)
	Leukaemia blood cells	(Depil *et al.* 2002)
	Germ cell and trophoblastic tumours	(Herbst *et al.* 1998)
HERV-R	Neoplastic tissues	(Andersson *et al.* 1998)
HERV-W	Ovarian carcinomas	(Menendez *et al.* 2004)
	Various cancer cell lines	(Yi *et al.* 2004)
LIN	Chronic myeloid leukaemia cell lines	(Roman-Gomez *et al.* 2005)
	Ovarian carcinomas	(Menendez *et al.* 2004)
	Urothelial carcinoma cell line	(Florl *et al.* 1999)

Table 1-2 Retroelements which have recently been associated with cancers.

The second exception relates to recent studies which have shown that endogenous non-telomerase reverse transcriptase may play a role in tumour growth (Sciamanna *et al.* 2005; Sinibaldi-Vallebona *et al.* 2006). Sciamanna and colleagues (2005) used RNAi directed towards ORF1 of intact L1 elements, to downregulate expression of these elements and therefore reduce the endogenous RT present in a cell. When A-375 cells were treated with RNAi to the L1 elements, they developed a differentiated morphology concomitant with reduced cell growth, which was comparable with cells treated with RT inhibitors. For cells to be expressing large amounts of RT from retroelements, it is likely that deregulation of the cell has already released the retroelements from their silent state; therefore these studies again suggest that the retroelements are not the direct cause of the disease state but rather a by-product of cell regulation breakdown.

1.7.6 Do retroelements cause disease?

Retroelement mediated genomic changes via insertions, deletions and unequal homologous recombination do cause diseases, but the majority of studies relating to retroelement expression and disease just look at transcript expression, and their studies at best only tenuously link retroelement expression with the diseased state. The studies do not show the retroelement to be the cause or even the reason for the disease to progress, but rather suggest the retroelements have become active as the cells are subject to deregulation due to the disease state. The studies do not even look at the possibility that the retroelements could be acting via transcriptional interference mechanisms to enhance the disease state, and no studies have asked if expression of a retroelement could be interfering with surrounding genes and be causing the diseased state.

1.8 Do retroelements have a biological function?

Retroelement activity is usually associated with negative connotations: the idea is that their activity is not part of the biology of the cell; therefore it can only be detrimental. There are many examples of retroelements transposing and causing disease, but there are also a number of papers which suggest that not all retroelements are detrimental to cell functioning and in fact some may have useful functions.

1.8.1 Retroelements as promoters, enhancers and polyadenylation signals

An obvious beneficial function of retroelements is when they have been appropriated by genes to act as alternate promoters, enhancers or stop sites of transcription for the gene (Yang *et al.* 1998; Mager *et al.* 1999; Baust *et al.* 2000; Landry *et al.* 2001; Dunn *et al.* 2003). ERV elements are good candidates for gene

appropriation due to the inbuilt promoter, enhancer and polyadenylation signals in their LTRs (refer to Figure 1-1).

In a study conducted by Medstrand *et al*. (2001) it was found that both the apolipoprotein C-I (apoC-I) gene and the endothelin B receptor (EBR) genes were expressed from alternate promoters provided by HERV-E LTRs. The EBR gene was only expressed from the alternate promoter in the placenta, but it was found to promote a significant proportion of the EBR transcripts. In contrast the apoC-I gene was expressed in many of the tissues tested, although it was preferentially expressed in the liver where approximately 15% of the transcripts were derived from the LTR promoter.

In another study conducted by Dunn and colleagues (2003) a HERV-L LTR was found to act as an alternate promoter for the β1,3-galactosyltransferase 5 gene in humans. It was shown to be the most active promoter (when compared to the native promoter and an additional alternate promoter) in the gastrointestinal tract and the mammary gland, as well as being the dominant promoter in the colon. Like the apoC-I and EBR genes, these results suggest retroelements may be regulated by external influences so they are expressed in a tissue specific manner.

ERV elements are not the only retroelements to be appropriated by genes. Yang *et al*. (1998) found a LINE element acts as an enhancer for the apolipoprotein(a) gene. The element confers a greater than 10 fold increase in activity in cultured hepatocyte cells, and it functions in either orientation. In this case the enhancer function of the retroelement for the apo(a) gene is beneficial, but it is obvious that the function could also easily be detrimental to a cell if a retroelement enhanced expression of a gene associated with disease.

Mager *et al*. (1999) and Baust *et al*. (2000) used similar approaches of searching cDNA libraries to identify gene transcripts which were polyadenylated by retroelement

LTRs. Mager and colleagues (1999) found that a HERV-H LTR provided the primary polyadenylation signal for the genes HHLA2 and HHLA3. No major non-LTR polyadenylation signals were detected for either gene, suggesting that the LTR had been appropriated a long time ago. Similarly Baust and colleagues (2000) found a HERV-K LTR that acts as a polyadenylation signal for an unknown gene, although they did not examine if a non-LTR polyadenylation signal was also present for the gene.

These are just a few examples where retroelements have been appropriated by genes for their use, and considering the large numbers of retroelements in the genome it is likely that there are more to be found. It is even possible that some of the transcripts we have identified in our experimental work are examples of such adapted use of retroelements.

1.8.2 Regulation of genes by retroelements

The ability of retroelements to regulate gene expression is not well understood, but studies suggest that in some cases it is occurring *in vivo*. The regulation seems to be most common at the transcription stage (Vidal *et al.* 1993; Peaston *et al.* 2004) but it has also been shown to occur at the translation stage (Landry *et al.* 2001).

Vidal and colleagues (1993) found that B1 elements were important for regulation of a complex group of mRNAs called piL genes. They were not sufficient for regulation but were shown to function as part of a sequence block, which upregulated the piL genes in the G_2 period of the first cell cycle, and increased their expression in transformed cells. Similarly Peaston and colleagues (2004) found MT elements, which were acting as alternative promoters and 5' exons for a subset of host genes, could be regulated by external influences resulting in differential expression in oocyte and preimplantation embryos.

A number of studies have found binding sites for transcription factors in various families of retroelements. ERV-9 elements have NFY, M2F1 and GATA-2 sites (Yu *et al.* 2005), L1 elements have RUNX3 and YY1 sites (Yang *et al.* 2003; Athanikar *et al.* 2004) and Alu elements have Sp1 and YY1 sites (Oei *et al.* 2004) just to give a few examples. In addition various retroelements have also been shown to be responsive to hormones, glucose, calcium, and other effectors (Leiter *et al.* 1986; Stavenhagen and Robins 1988; Schiff *et al.* 1991; McHaffie and Ralston 1995; Vansant and Reynolds 1995; Morales *et al.* 2002). It is therefore possible that co-ordinate regulation of some genes at the transcriptional level may be achieved through this mechanism, and in order to utilise these benefits the host has maintained the retroelements.

Retroelement-based regulation of translation has also been described by Landry *et al*. (2001). They examined repetitive elements in the 5' UTR of the zinc finger gene ZNF177 to see what affect they had on transcription and translation of the gene. When the Alu-L1 element region was placed into a reporter construct, it was shown to be capable of increasing transcription but it decreased translation efficiency. They speculated that this decrease was due to the possible secondary structure of the Alu element impeding ribosome binding. In a different study Rubin and colleagues (2002) also examined Alu elements and found they could stimulate translational expression by reducing the lag time between the end of transcription and the beginning of translation. They suggested this mechanism would be beneficial in cell stress recovery and viral infection, where the cell would be saved from wasting energy providing more transcripts by bringing forward the start of protein manufacture (Rubin *et al.* 2002). Results from the above studies suggest that the affect retroelements have on translation is driven by external factors, and that the cellular context of the gene determines if the affect will result in an upregulation or a downregulation of translation.

Further support for the retroelement-based gene regulation theory was found by Grover and colleagues (2003) who conducted an analysis of Alu distribution on human chromosomes 21 and 22 and found that their distribution was not random. Instead the elements were clustered in genes coding for metabolism, transport and signalling processes, and were poorly represented in genes coding for structural proteins and information pathway components. This was observed for both chromosomes even though Alu no. : Gene no. ratios differed for the two chromosomes. The authors concluded that this non-random distribution may reflect a positive selection of Alu elements within certain genes, due to their ability to regulate the genes (Grover *et al.* 2003).

1.8.3 Retroelements that have been adapted to become genes

Britten (2004) takes the idea of retroelement function one step further and claims that some coding sequences of functioning human genes are entirely composed of mobile element sequence. He has shown that the AD7C neuronal thread protein, BNIP3 gene involved in apoptosis control, and the syncytin gene which mediates trophoblastic fusion in the placenta, to name a few, are largely composed of retroelement sequence (Britten 2004).

A number of Ty3/gypsy type retrotransposons have also been found to be adapted as genes, although the new genes are not entirely composed of retroelement sequence as was observed by Britten. Ono and colleagues (2006) found evidence to suggest that the Peg10 imprinted gene in mice is derived from a member of the Ty3/gypsy retrotransposon group. Peg10 is thought to play a role in myelination, as it binds to the promoter of myelin basic protein. Not only is it a paternally imprinted gene, but it has also been shown to cause embryonic lethality when it is knocked out - indicating an essential role for a retrotransposon-derived gene (Ono *et al.* 2006). The

gene Rtl1, another Ty3/gypsy type retrotransposon like gene, has also been shown to be paternally imprinted and it has been suggested that miRNAs may play a role in the silencing of the gene (Seitz *et al.* 2003). This is interesting when it is noted that Youngson and colleagues (2005) found the Rtl1 gene to be methylated during development in a similar way to other retroelements. Taken together it seems that the retroelement type silencing has been adapted to control the expression of the Rtl1 gene (Youngson *et al.* 2005).

There are likely to be other examples where retroelements have degenerated and have been adapted in full as genes; these are clear examples of the biological significance of retroelements.

1.8.4 Retroelement-driven genomic changes

Genomic changes induced by retroelements have also been suggested to have beneficial properties. RNA editing, alternate splicing, immunological diversity, nucleosome positioning and genomic expansion (Yang and Dudley 1992; Kidwell and Lisch 1997; Kulski *et al.* 1997; Sorek *et al.* 2002; Lev-Maor *et al.* 2003; Liu *et al.* 2003; Kim *et al.* 2004; Zheng *et al.* 2005) are just some ways retroelements have been shown to contribute to the diversity of a genome.

RNA Editing

RNA editing by ADAR (adenosine deaminase that act on RNA) enzymes results in the site-specific conversion of A-I bases in precursor mRNA messages (Grover *et al.* 2005). This base change destabilises double stranded RNA structures by converting AU base pairs to unstable IU base pairs, and it is thought to antagonise the RNAi pathway (Scadden and Smith 2001). A study conducted by Kim and colleagues (2004) showed that the majority of A-I editing occurs within transcribed sense or antisense Alu sequences. 2% of all publicly available full length cDNAs from >250 human libraries

contained edited Alu sequences. The edited Alu sequences were primarily associated with retained introns, extended UTRs, or with transcripts that have no corresponding known gene, indicating the editing may serve as a mark for transcripts which are not destined for translation.

Alternate Splicing

Alu elements have also been implicated in the process of exonization whereby Alu elements are inserted into mature mRNAs via a splicing-mediated process (Lev-Maor *et al.* 2003). Sorek and colleagues (2002) found that more than 5% of human alternatively spliced exons were Alu derived. This is a possible benefit for the cell, as the original host protein is still produced, but the cell has the opportunity to provide function to the new alternatively spliced protein which is also being manufactured. In a later study the same lab found that a mutation preceding the proximal of two possible splice sites within the Alu element could cause the alternate splicing event to become constitutive. This type of exonization is likely to have detrimental affects as the host protein is no longer being formed (Lev-Maor *et al.* 2003), suggesting exonization can be a double edged sword. An additional study by Zheng *et al.* (2005) found evidence via computational methods that LINE elements are also associated with alternate splicing although ERV elements are not. Therefore the LINE elements, like the Alu elements, may also provide new genetic material via splicing mechanisms.

Immunological Diversity

Retroelement mobility via transposition is thought to have some beneficial consequences, not just the negative consequences of insertional mutations, one of which is the ability to provide diversity to immunological genes (refer to Figure 1-9 for transposition examples). 20% of the MHC class II region is composed of retroelements

(Andersson *et al.* 1998) and it has been proposed that different retroelement insertions contribute to MHC haplotype polymorphisms.

A study conducted by Kulski and colleagues (1997) found that the HLA-B and HLA-C genes are contained within the duplicated segments peri-B and peri-C, and the duplicated segments share 90% homology with the parent genes except when interrupted by retroelement insertions or deletions (Kulski *et al.* 1997). These retroelements serve to diversify the region and maintain the plasticity of the region.

It is also possible that retroelements act to regulate the region as discussed in Section 1.8.2. ERV-9 elements, which are found within the MHC region, have been shown to have interferon-γ responsive elements (γ-IRE). This is of interest as IFN-γ is a potent inducer of HLA-DR gene expression, and it has therefore been speculated that the LTRs of the ERV-9 elements may act as transcriptional enhancers for HLA-DR genes (Andersson *et al.* 1998).

A more direct action of retroelements on the immune system was suggested by Yang and Dudley (1992) who observed that the retroelement Mtv-8 (an MMTV retroelement) appeared to enhance $V_{\kappa}9M$ gene rearrangements. Their observations were in part based on the fact that $V_{\kappa}9M$ rearrangements were 5-10 fold higher in spleens from Mtv-8 positive mice, compared to spleens from mice which lacked the retroelement insert.

Diversity within the immune system is highly beneficial for survival so it is an interesting concept that retroelements may in fact play a part in maintaining a healthy immune system rather than being one of its main targets.

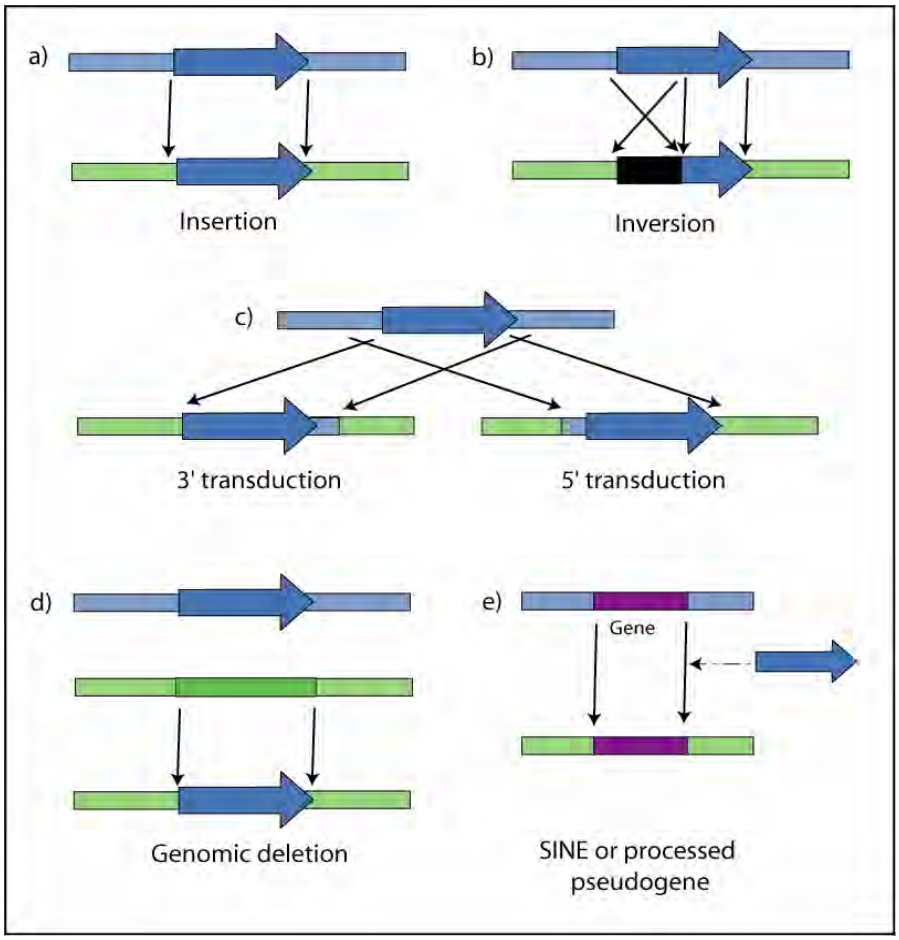

Figure 1-9 Examples of genomic rearrangements due to transposition of retroelements. Retroelements are shown as blue arrows, genes as a purple line, the original genomic location is shown as a blue line and the new genomic location is shown as a green line. **a)** An insertion is where a copy of a retroelement is inserted into a new locus disrupting the locus but not deleting any of the sequence from the locus. **b)** An inversion is where half of the retroelement is inserted in the reverse orientation causing the disruption of both the retroelement and the new locus. **c)** Transduction is where the retroelement mobilises flanking sequence which is 5' or 3' to itself and this sequence is inserted elsewhere in the genome with the retroelement sequence. **d)** A genomic deletion occurs when the retroelement copy replaces sequence from the new genomic locus. **e)** SINEs and processed pseudogenes are mobilised when other retroelements such as LINEs provide enzymatic machinery for them to transpose to a new genomic location. Adapted from (Deininger *et al.* 2003)

Nucleosome Positioning

In order to fit a large genome into a nucleus, eukaryotic chromatin "packs" the DNA; the primary units of packing are DNA histone complexes called nucleosomes. It is thought that nucleosomes are positioned at specific genomic sites via functional patterns of genomic sequences (Grover *et al.* 2005). Alu elements have been shown to direct translational and rotational positioning of octamer as well as tetramer nucleosome particles in *in vitro* experiments. The positioning signal occurs between the A and B box of the RNA polymerase III promoter motifs in the most highly conserved region of the Alu sequence (Englander *et al.* 1993). These results were confirmed by *in vivo* studies which found that Dnase I treatment of the region spanning the RNA polymerase III Box A promoter region and upstream sequence produced a pattern of nicks suggesting that Alus influence nucleosome formation over neighbouring regions (Englander and Howard 1995). The group then used a computational analysis to show that the information for nucleosome positioning is actually intrinsic to Alu elements. These results taken together with the large numbers of Alu elements in the genome, suggest Alus could be playing a role in chromatin organisation by positioning of nucleosomes over certain regions.

Genomic Expansion

Through their method of replication, retroelements have amplified in number resulting in an expansion of the mammalian genome over time. In the last 50 million years the human genome has expanded by 15-20%, with 90% of this expansion being due to retroelement insertions (Liu *et al.* 2003). The maintenance of these retroelements is thought to increase the plasticity of the mammalian genome thereby leaving room for genome evolution. By providing variation retroelements may be adapted by the host, as mentioned in the above subchapters, to cope with stress and detrimental conditions

(Kidwell and Lisch 1997). The variation is most commonly attributed to the transposition affects such as insertional mutations, exon shuffling, gene breaking and the production of pseudogenes as well as affects such as recombination and exonization.

Nevertheless little thought is given to transcriptional interference, which is also capable of driving variation: the means are not overtly obvious, but may be equally as potent as genetic mutations.

1.9 The retroelement dilemma

Our lab is interested in the ability of retroelements to exert transcriptional interference on surrounding genes, but working with these elements presents a problem. Retroelements are defined as repetitive elements, which means they are present in numerous copies in the genome (see Table 1-1). This poses the obvious dilemma of how to examine a single retroelement when its location is unknown. Probes designed for Northern and Southern blots will be guaranteed to bind to more than one retroelement from any given family, while PCR primers are equally guaranteed of amplifying more than just the retroelement of interest. Large scale microarrays are also out of the question, due to problems designing a specific probe (unless flanking sequence is known). This has been the problem encountered by groups studying retroelements. It is still a problem, although with the sequencing of the human and mouse genome, added to the better range of BLAST programs, some progress can be made. Amplified fragment length polymorphism (AFLP) based techniques have also recently been shown to be useful for looking at a range of active retroelements (Vinogradova *et al.* 2002; Kashkush *et al.* 2003), but on the whole techniques used for studying retroelements are limited.

1.10 Aims of this thesis

Various classes of retroelements have been shown to be active and capable of transcription, but no large scale study has been conducted to look at all families of retroelements to see how many have intact promoters that are capable of transcription. We are interested in the ability of these retroelements to be transcriptionally active, as transcriptional interference may occur commonly in both mice and humans. A better understanding of the extent of this phenomenon may have an impact on our views of the origin of phenotypic differences between individuals and variation in disease risk.

This thesis aims to -

1. Use a computational approach to examine the EST database and estimate how many retroelements are capable of transcription in humans and mice, and investigate where the active retroelements are located in the genomes, in order to estimate relative numbers of unique active elements.
2. Gain experimental evidence that retroelements in the mouse are capable of exerting transcriptional interference on surrounding genes; this study uses an AFLP-based method.
3. Provide evidence for the authenticity of retroelement transcripts that are candidates for causing transcriptional interference, by using primer extension and RT-PCR.

CHAPTER 2 MATERIALS AND METHODS

2.1 General Materials

2.1.1 General materials and reagents

Analytical grade organic solvents such as ethanol, isoamyl alcohol and propan-2-ol were purchased from BDH Laboratory Supplies (England). Trizol, formamide and phenol were purchased from Invitrogen (USA).

Ampicillin, Actinomycin D and diethyl-pyrocarbonate (DEPC) were obtained from Sigma (USA). Isopropyl β-D-thiogalactosidase (IPTG) and 5-bromo-4-chloro-β-D-galactopyranoside (X-gal) were purchased from Promega (USA).

N,N,N',N'-tetra-methyl-ethylenediamine (TEMED), 40% acrylamide/bis-acrylamide solution, ammonium persulfate, ethidium bromide, xylene cyanol and electrophoresis purity bromophenol blue were obtained from Bio-Rad Laboratories (USA). Deoxynucleotides were purchased from Roche Applied Science (Germany). Urea was purchased from Invitrogen (USA). Sheared herring sperm DNA was purchased from Promega (USA). Sigmacote and 1,4-Dithio-DL-threitol were purchased from Sigma/Aldrich (USA). DNA grade agarose and low melt agarose were purchased from Progen Biosciences (Australia). BigDye terminator RR mix was purchased from Applied Biosystems (USA).

Biomax imaging film was purchased from Kodak (USA).

The following solutions were obtained in house - 50 x TAE, 50 x TBE, Luria-Burtani (LB) medium, LB agar, 10 x TE, 0.5M EDTA, 1M $MgCl_2$, 3M KCl, 3M sodium acetate pH 5.0, 5M NaCl and 1M Tris pH 8.0.

2.1.2 Buffers and Solutions

Buffer TAE consists of 40mM Tris pH 8.0, 1% glacial acetic acid and 5mM EDTA. Buffer TBE is composed of 40 mM Tris pH 8.0, 100mM boric acid and 5mM EDTA. Buffer TE consists of 10mM Tris pH 8.0, 1mM EDTA.

The Transposon display running buffer was made using 98% formamide, 0.01 M EDTA, 0.005% Bromophenol blue and 0.005% Xylene cyanol.

2.1.3 Enzymes

The following restriction endonucleases from NEB (USA) were used in this thesis; Acc I, Aci I, ApaL I, BsiHKA I, Eco0109 I, EcoR I, Hha I, Hind III, Hpa II, HpyCH4 IV, Mse I, Sty I, Xma I. The enzymes DNA Polymerase I (*E. coli*), *E. coli* DNA Ligase, T4 DNA Ligase, T4 DNA Polymerase and T4 Polynucleotide Kinase were also purchased from NEB (USA). Recombinant RNasin Ribonuclease Inhibitor and Rq1 RNase-Free Dnase were purchased from Promega (USA). RNase, DNase-free and Taq DNA Polymerase were purchased from Roche Applied Science (Germany). SuperScript II Reverse Transcriptase was purchased from Invitrogen (USA).

2.1.4 Radiochemicals

[γ-^{32}P]-ATP (3000 Ci/mmol) aqueous solutions were purchased from Amersham (UK).

2.2 Computer Analysis

2.2.1 EST database search for Human ERV retroelements

Regions surrounding R from the following human ERV retrotransposon sequences were used as probes for the database search; ERV-1 (K02919 1-480 bp), ERV-9 (AF064191 421-982 bp), HERV-E (M10976 251-650 bp), HERV-F (Kjellman *et al.* 1999 5′ LTR 301-536 bp), HERV-H (AF108838 301-700 bp), HERV-L (X89211

1-462 bp), HERV-R (K02016 361-618 bp), HERV-W (AF127226 201-600 bp) and HTLV-1 (X16660 1-684 bp). Additional probes were taken from Repbase (http://www.girinst.org/index.html) release 8.1.5; HERV-I (LTR10A 301-609 bp), HERV-P (LTR8 401-691 bp), RRHERV-I (LTR15 241-493 bp) and S71 (LTR6B 301-558 bp). The probes were used to BLASTn search the NCBI Human EST database (http://www.ncbi.nlm.nih.gov) release 122 - 135 using a maximum limit of 1000 alignments. Resulting ESTs were included for further analysis if the transcript initiated or terminated in the R region and the LTR region was highly-moderately conserved. A search of the dbEST was then conducted using the EST clone IDs to determine if the other end of the clone had been sequenced. Finally the ESTs were used to BLAST the Human genome (filter removed) to determine the chromosomal location of the EST. The surrounding chromosomal sequence was entered into Repeatmasker (http://woody.embl-heidelberg.de/repeatmask) to show whether the EST was solo or full length. All searches were conducted using the NCBI website, unless otherwise specified, and they were carried out on a iBook personal laptop.

2.2.2 EST database search for Human SINE/LINE retroelements

Probes for the human LINE and SINE elements were taken from Repbase release 8.1.5; the promoter region of the LINE element was used (L1 consensus 1-1030 bp) and a full length SINE (Alu consensus 1-290 bp). The probes were used to search the Genbank Human EST database as above. Resulting LINE ESTs were included for further analysis if the transcripts were initiated or terminated at a location within the 5′ UTR and the promoter region was highly-moderately conserved. Resulting SINE transcripts were included if the transcripts initiated or terminated at the very 5′ or 3′ end of the probe. The SINE and LINE ESTs were then analysed as above to determine if

both ends of the clone had been sequenced and to find the chromosomal location of the ESTs.

2.2.3 EST database search for Mouse ERV retroelements

Regions surrounding R from the following mouse ERV retrotransposon sequences were used as probes for the database search; ETn (X03063 1-322 bp), GLN (M14005 (1-618 bp), A^{vy} IAP (Duhl *et al.* 1994 1-337), MeRV-L (Y12713 250-530 bp), MMTV (AF228550 900-1480 bp), MuERV-C (AF049340 5661-6234 bp), MuRRS (X02487 1-509 bp), MuRVY (M27506 (1-627 bp), MuLV (X05157 480-720 bp) and VL30 (M21123 200-500 bp). The methods used for the database searches are as described in Section 2.2.1 however the mouse EST and Mouse genome databases were searched.

2.2.4 EST database search for Mouse SINE/LINE retroelements

Probes for the LINE element database search were taken from the three main active LINE families in the mouse; L1A (M29324 1092-1837 bp), L1Gf (AC068252 161291-159839 bp) and L1Tf (Naas *et al.* 1998 1-212 bp). SINE probes were taken from Repbase release 8.1.5; B1 (1-135 bp), B2 (1-209 bp) and B3 (1-226 bp). Retrieved sequences were analysed as described in Section 2.2.2 with the exception that mouse databases were used instead of human.

2.3 Transposon Display

2.3.1 RNA Extraction

RNA was extracted from the liver or kidney of either an a/a C57Bl/6 mouse, an A^{vy}/a C57Bl/6 mottled mouse or an A^{vy}/a C57Bl/6 yellow mouse. All mice were obtained from the Garvan Institute Animal Facility. Both male and female mice were used and their ages ranged from 3 weeks to 9 months. Tissue choice was random and

was driven by the fact that the liver and kidneys could provide a large amount of RNA for the experiments.

The tissue was homogenized in Trizol reagent and total RNA was extracted following the protocol provided with the Trizol reagent. 50 µl of RNA (100-1000 µg) was then DNase treated with 20 U of Rq1 RNase-Free Dnase and 160U of RNasin Ribonuclease Inhibitor in the provided buffer, for 30mins at 37°C before being cleaned up on Phase Lock tubes (Eppendorf, Germany) following the provided protocol. RNA was stored at -80°C until needed.

2.3.2 Quantification and Analysis

RNA was quantified using a SmartSpec 3000 spectrophotometer (BioRad, USA) to calculate absorbance at 260 nm and this was applied to the equation: A 1.0_{260}=40 µg/ml RNA. Approximately 1 µg of RNA was then run at 110 V for approximately 30 min on a 1% agarose TAE gel, with 0.3 µg/ml of ethidium bromide, to check the quality of the extracted RNA. The GelDoc 2000 (BioRad, USA) Quantity 1 – 4.1.0 software was used to visualize and record the image.

2.3.3 cDNA Synthesis

Double stranded cDNA was synthesized using the SuperScript Double-Stranded cDNA Synthesis Kit from Invitrogen. 100 pmol of oligodT primer (Roche Applied Science, Germany) was added to 25 µg of total RNA (total RNA was extracted from tissue taken from a single mouse ie 1 a/a C57Bl/6 mouse, 1 A^{vy}/a C57Bl/6 mottled mouse and 1 A^{vy}/a C57Bl/6 yellow mouse) for the reaction and the resultant cDNA was resuspended in 40 µl of water. 10 µl of cDNA was then run for approximately 40 min at 110 V on a 1.5% agarose TAE gel, with 0.3 µg/ml ethidium bromide, so the quality of cDNA could be observed.

2.3.4 Transposon Display Oligonucleotides

Oligonucleotides which were used in the transposon display experiments are shown in Table 2-1-Table 2-6. They were supplied by either Sigma-Genosys (Australia) or Geneworks (Australia). The IAP oligonucleotides were designed to the sequence taken from Duhl *et al.* (1994). Antisense LINE oligonucleotides were designed to the sequence taken from Naas *et al.* (1998) and sense oligonucleotides were designed to the sequence Genbank Accession AF081113. The MuLV oligonucleotides were designed to Genbank Accession X05157 while the MMTV oligonucleotides were designed to Genbank Accession AF228550. The SINE oligonucleotides were designed to the B1 sequence taken from Repbase release 8.1.5 while the VL30 oligonucleotides were designed to Genbank Accession M21123.

IAP Adaptor And Primer Oligonucleotides	Supplier	Sequence
IAPHae/ad1	Geneworks	5' CATACACGGAAC 3'
IAPHae/ad2	Geneworks	5' GGATGTTCCGTGTATGGCGC 3'
IAPHind/ad1	Sigma Genosys	5' CGAGCGTAGTGCTAGG 3'
IAPHind/ad2	Sigma Genosys	5' AGCTCCTAGCACTACG 3'
H-HTaq/ad3	Geneworks	5' CCGATGAGCGTGCACTT 3'
H-HTaq/ad4	Geneworks	5' CGAAGTGCACGCTCA 3'
IAPHaepre1	Geneworks	5' GTTCCGTGTATGGCGCCATC 3'
IAPHindpre1	Sigma Genosys	5' CGTAGTGCTAGGAGCTTTGTC 3'
H-HTaqpre2	Geneworks	5' CCGATGAGCGTGCACTTCGA 3'
IAPintamp1	Geneworks	5' GTGACGGCGAATGTGGGGG 3'
IAPintamp2/3	Sigma Genosys	5' TCCTGGCCGGTCGGTGAAG 3'
H-HTaqamp5	Geneworks	5' CCGATGAGCGTGCACTTCGANNNA 3'
H-HTaqamp6	Geneworks	5' CCGATGAGCGTGCACTTCGANNNT 3'
H-HTaqamp7	Geneworks	5' CCGATGAGCGTGCACTTCGANNNC 3'
H-HTaqamp8	Geneworks	5' CCGATGAGCGTGCACTTCGANNNG 3'

Table 2-1 Oligonucleotides used in the IAP Transposon Display

LINE Adaptor And Primer Oligonucleotides	Supplier	Sequence
L1TfEco0/ad1	Sigma Genosys	5' GACTGAATCAGCAC 3'
L1TfEco0/ad2	Sigma Genosys	5' CTAGGTGCTGATTCA 3'
L1TfBsi/ad1	Sigma Genosys	5' CGATGTTGACGACAGGTAGCT 3'
L1TfBsi/ad2	Sigma Genosys	5' ACCTGTCGTCAAC 3'
H-HTaq/ad1	Sigma Genosys	5' CGACCTGATCACGGA 3'
H-HTaq/ad2	Sigma Genosys	5' CAGGTCCGTGATCAGGT 3'
L1TfEco0pre1	Geneworks	5' GTGCTGATTCAGTCCCGGAC 3'
L1TfBsipre1	Geneworks	5' GTTGACGACAGGTAGCTCTTG 3'
H-HTaqpre1	Geneworks	5' CAGGTCCGTGATCAGGTCGA 3'
L1Tfintamp1	Sigma Genosys	5' CTGTGGCTTAGGCCGCCTC 3'
L1Tfintamp2/3	Sigma Genosys	5' CGGCCATCACTGGAAAGAGAG 3'
H-HTaqamp1	Geneworks	5' CAGGTCCGTGATCAGGTCGANNNA 3'
H-HTaqamp2	Geneworks	5' CAGGTCCGTGATCAGGTCGANNNT 3'
H-HTaqamp3	Geneworks	5' CAGGTCCGTGATCAGGTCGANNNC 3'
H-HTaqamp4	Geneworks	5' CAGGTCCGTGATCAGGTCGANNNG 3'

Table 2-2 Oligonucleotides used in the LINE Transposon Display

MMTV Adaptor And Primer Oligonucleotides	Supplier	Sequence
MMTVApaL/ad1	Geneworks	5' TGCATGGCATACTGC 3'
MMTVApaL/ad2	Geneworks	5' CTAGGCAGTATGCCA 3'
MMTVHae/ad1	Geneworks	5' GCTCCTTGATACTATGCGC 3'
MMTVHae/ad2	Geneworks	5' ATAGTATCAAG 3'
MMTVHpa/ad1	Geneworks	5' CGATCGCTAAGCATCAGATG 3'
MMTVHpa/ad2	Geneworks	5' CGGACATCTGATGCTTAGCGAT 3'
MMTVHpa/ad3	Geneworks	5' CTGAGCGTTCCAGTGCAAGAT 3'
MMTVHpa/ad4	Geneworks	5' CGATCTTGCACTGGAACGC 3'
MMTVApaLpre1	Geneworks	5' GCAGTATGCCATGCACGCAG 3'
MMTVHaepre1	Geneworks	5' CTTGATACTATGCGCCCGAAC 3'
MMTVHpapre1	Geneworks	5' CATCTGATGCTTAGCGATCGG 3'
MMTVHpapre2	Geneworks	5' GCGTTCCAGTGCAAGATCGG 3'
MMTVintamp1/3	Sigma Genosys	5' CTTGGCTGCTTCTCTCCTAAG 3'
MMTVintamp2	Geneworks	5' CGGATAAGTGACCCTTGTCTC 3'
MMTVHpaamp1	Geneworks	5' CTGATGCTTAGCGATCGGNNNA 3'
MMTVHpaamp2	Geneworks	5' CTGATGCTTAGCGATCGGNNNC 3'
MMTVHpaamp3	Geneworks	5' CTGATGCTTAGCGATCGGNNNG 3'
MMTVHpaamp4	Geneworks	5' CTGATGCTTAGCGATCGGNNNT 3'
MMTVHpaamp5	Geneworks	5' GTTCCAGTGCAAGATCGGNNNA 3'
MMTVHpaamp6	Geneworks	5' GTTCCAGTGCAAGATCGGNNNC 3'
MMTVHpaamp7	Geneworks	5' GTTCCAGTGCAAGATCGGNNNG 3'
MMTVHpaamp8	Geneworks	5' GTTCCAGTGCAAGATCGGNNNT 3'

Table 2-3 Oligonucleotides used in the MMTV Transposon Display

MuLV Adaptor And Primer Oligonucleotides	Supplier	Sequence
MuLVApaL/ad1	Geneworks	5' TGCATGCTAACGCAGG 3'
MuLVApaL/ad2	Geneworks	5' CTAGCCTGCGTTAGCA 3'
MuLVSty/ad1	Sigma Genosys	5' GGATGATACAGGACTCA 3'
MuLVSty/ad2	Sigma Genosys	5' CAAGTGAGTCCTGTATC 3'
MuLVAci/ad1	Sigma Genosys	5' GACGCTGAGGTACATTGCTGTA 3'
MuLVAci/ad2	Sigma Genosys	5' CGTCAGCAATGTACCTCAG 3'
MuLVApaLpre1	Geneworks	5' CCTGCGTTAGCATGCACCTC 3'
MuLVStypre1	Sigma Genosys	5' GATACAGGACTCACTTGGGAG 3'
MMTVApaLpre1	Geneworks	5' GCAGTATGCCATGCACGCAG 3'
MuLVAcipre1	Sigma Genosys	5' CTGAGGTACATTGCTGTACGC 3'
MuLVintamp1/3	Sigma Genosys	5' CTTGGCTGCTTCTCTCCTAAG 3'
MuLVintamp2/2	Sigma Genosys	5' GGTCTCCTCAGATTGATTGAC 3'
MuLVAciamp1	Sigma Genosys	5' GGTACATTGCTGTACGCNNNT 3'
MuLVAciamp2	Sigma Genosys	5' GGTACATTGCTGTACGCNNNA 3'
MuLVAciamp3	Sigma Genosys	5' GGTACATTGCTGTACGCNNNC 3'
MuLVAciamp4	Sigma Genosys	5' GGTACATTGCTGTACGCNNNG 3'

Table 2-4 Oligonucleotides used in the MuLV Transposon Display

SINE Adaptor And Primer Oligonucleotides	Supplier	Sequence
B1Acc/ad1	Geneworks	5' GCGACCTTCAGCGTTAC 3'
B1Acc/ad2	Geneworks	5' AGGTAACGCTGAAGG 3'
B1Acc/ad3	Geneworks	5' CTGACACCTCAAGTAG 3'
B1Acc/ad4	Geneworks	5' CTCGCTACTTGAGGTGTC 3'
B1HpyC/ad1	Geneworks	5' CGGTTCAGAACTGGATCAGC 3'
B1HpyC/ad2	Geneworks	5' CCATGCTGATCCAGTTCTGAAC 3'
B1HpyC/ad3	Geneworks	5' CTCAGTCCTGACGCTTCGATC 3'
B1HpyC/ad4	Geneworks	5' CGGATCGAAGCGTCAGGAC 3'
B1Accpre1	Geneworks	5' CCTTCAGCGTTACCTACAGAG 3'
B1Accpre3	Geneworks	5' CTACTTGAGGTGTCAGACCAG 3'
B1HpyCpre1	Geneworks	5' GCTGATCCAGTTCTGAACCGT 3'
B1HpyCpre2	Geneworks	5' GTCCTGACGCTTCGATCCGT 3'
B1intamp1	Geneworks	5' CTGCCTCCCAAGTGCTGGG 3'
B1intamp2	Geneworks	5' CCAGGACAGCCAGGGCTAC 3'
B1HpyCamp1	Geneworks	5' GATCCAGTTCTGAACCGTNNNA 3'
B1HpyCamp2	Geneworks	5' GATCCAGTTCTGAACCGTNNNC 3'
B1HpyCamp3	Geneworks	5' GATCCAGTTCTGAACCGTNNNG 3'
B1HpyCamp4	Geneworks	5' GATCCAGTTCTGAACCGTNNNT 3'
B1HpyCamp5	Geneworks	5' CTGACGCTTCGATCCGTNNNA 3'
B1HpyCamp6	Geneworks	5' CTGACGCTTCGATCCGTNNNC 3'
B1HpyCamp7	Geneworks	5' CTGACGCTTCGATCCGTNNNG 3'
B1HpyCamp8	Geneworks	5' CTGACGCTTCGATCCGTNNNT 3'

Table 2-5 Oligonucleotides used in the SINE Transposon Display

VL30 Adaptor And Primer Oligonucleotides	Supplier	Sequence
VL30Sty/ad1	Sigma Genosys	5' CAAGTTCGTATGGTCATC 3'
VL30Sty/ad2	Sigma Genosys	5' GCGTGATGACCATACGA 3'
VL30Xma/ad1	Geneworks	5' GCGAGTCATAGCAAT 3'
VL30Xma/ad2	Geneworks	5' CCGGATTGCTATGAC 3'
VL30Hha/ad1	Geneworks	5' TAGGTCATCCGACGAAG 3'
VL30Hha/ad2	Geneworks	5' GTTCCTTCGTCGGATGACCTACG 3'
SINEMse/ad3	Geneworks	5' CCTGGATGACCAGTTGCG 3'
SINEMse/ad4	Geneworks	5' TACGCAACTGGTCATC 3'
VL30Stypre1	Sigma Genosys	5' GATGACCATACGACTTGGGAC 3'
VL30Xmapre1	Geneworks	5' GTCATAGCAATCCGGGACTTG 3'
VL30Hhapre1	Geneworks	5' CTTCGTCGGATGACCTACGC 3'
SINEMsepre2	Geneworks	5' CCTGGATGACCAGTTGCGTAA 3'
VL30intamp1/3	Sigma Genosys	5' GGTTCTGCCAAAGGATTCTACG 3'
VL30intamp2	Sigma Genosys	5' GTCTGAGTGAGGGTCTTCCC 3'
VL30Hhaamp1	Geneworks	5' CGTCGGATGACCTACGCNNNA 3'
VL30Hhaamp2	Geneworks	5' CGTCGGATGACCTACGCNNNC 3'
VL30Hhaamp3	Geneworks	5' CGTCGGATGACCTACGCNNNG 3'
VL30Hhaamp4	Geneworks	5' CGTCGGATGACCTACGCNNNT 3'
SINE Mseamp1	Geneworks	5' GGATGACCAGTTGCGTAANNNC 3'
SINE Mseamp2	Geneworks	5' GGATGACCAGTTGCGTAANNNG 3'
SINE Mseamp3	Geneworks	5' GGATGACCAGTTGCGTAANNNA 3'
SINE Mseamp4	Geneworks	5' GGATGACCAGTTGCGTAANNNT 3'

Table 2-6 Oligonucleotides used in the VL30 Transposon Display

2.3.5 Radiolabelling of Oligonucleotides for Transposon Display

The oligonucleotide primers were labelled following the protocol outlined by Triezenberg (1992). Labelled primers were purified using Micro BioSpin 6 columns (BioRad, USA) following the provided protocol.

2.3.6 Transposon Display method

Transposon displays were carried out as described by Kashkush *et al.* (2002) with some variations. Refer to Figure 4-1 for a pictorial representation of the method.

The initial step involves finding a rare base cutter (such as a penta or hexa cutter), which will cut the retroelement of interest in only one place, and a tetracutter which will cut unknown sequence downstream or upstream of the retroelement but will

not cut the retroelement. Refer to Table 2-7 for the restriction enzymes used in this study.

RETROTRANSPOSON FAMILY	ANTISENSE RESTRICTION ENZYMES	SENSE RESTRICTION ENZYMES
IAP	Hae II, Taq I	Hind III, HpyCH4 IV
LINE	Eco0109 I, Taq I	BsiHKA I, Taq I
MMTV	ApaL I, Hpa II	Hae II, Hpa II
MuLV	ApaL I, Aci I	Sty I, Taq I
SINE	Acc I, HpyCH4 IV	Acc I, HpyCH4 IV
VL30	Sty I, Mse I	Xma I, Hha I

Table 2-7 Restriction enzymes used for the mouse transposon display. Restriction enzymes are displayed for the mouse retroelement families examined for both the antisense and sense direction reactions. The rare base cutter is listed first followed by the tetracutter for both the antisense and sense reactions.

Approximately 300 ng of cDNA was digested with 1.2 U of the tetracutter and 10 U of the rare base cutter in a 50 µl final volume. Either a double digest or a sequential digest was conducted depending on the enzyme temperature compatibility. Adaptor sequences, designed so the restriction site is not conserved, were ligated to both ends of the fragment by adding 120 U of T4 ligase to the mix. Refer to Table 2-1-Table 2-6 for adaptor sequences. The samples were left for 1 hr at room temperature before an additional restriction step (as above) was conducted so any ligation artefacts were removed.

A preamplification round of PCR was then carried out using primers which are designed within the adaptor sequence for the tetracutter adaptor and partially within the adaptor/partially within the retroelement for the rare base cutter adaptor. 100 ng of each primer was added to approximately 50 ng of the restriction/ligation samples in a PCR mix consisting of a 10 mM Tris-HCl, 50 mM KCl pH 8.3 buffer, 2.5 mM $MgCl_2$, 0.1 mM dNTPs and 1 U of Taq polymerase in a 20 µl final volume. The PCR was run for

19 cycles with a 30 sec 94°C denaturation step, a 1 min 56°C annealing step and a 1 min 72°C extension step. Once the reaction was complete 10 µl of the PCR reaction was run on a 1.5% agarose TAE gel at 110 V for approximately 30 min, with 0.3 µg/ml of ethidium bromide, to check the preamplification had worked. The GelDoc 2000 (BioRad, USA) Quantity 1 – 4.1.0 software was used to visualize and record the image. The remaining 10 µl of PCR reaction was diluted with 190 µl of water.

An amplification round of PCR was then conducted using a [γ-^{32}P]-ATP labelled primer designed to sequence within the retroelement and priming towards the retroelement's adjacent genomic sequence, and a primer designed to the adaptor sequence with an additional 4 random nucleotides at the 3' end. 3 µl of dilute preamplification sample was added to 100ng of each primer in a mix as described above. A touchdown PCR was then conducted by running one cycle with a 30 sec 94°C denaturation step, a 30 sec 70°C annealing step and a 2 min 72°C extension step followed by 15 cycles where the annealing temperature was lowered by 1°C each cycle. The annealing temperature was then kept at 55°C for 35 cycles so products could be amplified.

The radioactive samples were then run on a 5% denaturing polyacrylamide sequencing gel so bands could be visualised. The gel was preheated before the samples were loaded by running it in 1x TBE at 2000V until the glass plates were approximately 56°C. The samples were heated at 96°C for approximately 5 min before 6-8 µl was loaded on the gel. After electrophoresis the gel was dried on a slab gel drier and it was then exposed to film overnight at -80°C.

2.3.7 Reamplification Of Transposon Display Bands

Bands of interest were excised from the gel, placed in approximately 300 µl of water (the amount varied depending on the band size) and left at 55°C for 2 hr. Since

our samples were run in quadruplicate, all four bands were placed in the same tube to increase the yield of DNA. The samples were then syringe filtered using glass wool before being cleaned up with a QIAEX II gel extraction kit (QIAGEN, Germany), following the provided protocol, to remove excess urea. 10 µl of each sample along with 100 ng of the amplification primer designed within the retroelement (intamp primers) and 100 ng of the preamplification primer designed within the tetracutter adaptor, was then added to a PCR mix consisting of a 10 mM Tris-HCl, 50 mM KCl pH 8.3 buffer, 2.5 mM $MgCl_2$, 0.05 mM dNTPs and 1 U of Taq polymerase in a 50 µl final volume. The PCR was run for 50 cycles with a 30 sec 94°C denaturation step, a 1 min 56°C annealing step and a 1 min 72°C extension step.

The PCR products were run on 1% low melt agarose TAE gels with 0.3 µg/ml of ethidium bromide. They were run at 110 V for approximately 30 min before being visualised on the GelDoc 2000 (BioRad, USA) using the Quantity 1 – 4.1.0 software. Resulting bands were excised and cleaned up using QIAquick PCR Purification kits (QIAGEN, Germany). 5 µl of each sample was then run on a 1% agarose gel (as described above) to approximate the concentration.

2.3.8 Cloning Of Transposon Display Bands

Samples were ligated into the pGemTEasy plasmid (Promega, USA) using 25 ng of plasmid, 400 U T_4 ligase in a 50 mM Tris-HCl (pH 7.5), 10 mM $MgCl_2$, 10 mM dithiothreitol, 1 mM ATP, 25 µg/ml bovine serum albumin buffer and between 5-15 µl of sample (depending on the approximate concentration) in a final volume of 20 µl. The samples were left at room temperature for 30 min before being transformed. Transformations were carried out into DH5α chemically competent cells following the pGemT transformation protocol (Promega, USA) - with the exception that the incubation step was conducted for 15 min without shaking. Cells were plated onto LB

plates with 100 mg/ml ampicillin plus 800 µg of X-gal and 4 µl of 1 M IPTG. Positive colonies were mini prepped using the FastPlasmid Mini kits (Eppendorf, Germany) and protocol provided.

2.3.9 DNA Sequencing

Reamplified products were sequenced following the protocol provided by the UNSW Sequencing facility. The M13 -21 primer (Applied Biosystems, USA) was used for sequencing.

2.3.10 Transposon Display Sequence Analysis

Sequences for the excised bands were analysed using the DNAStar alignment software as well as the NCBI Human and Mouse Genome BLAST programs.

2.4 Primer Extension

2.4.1 Primer Extension Oligonucleotides

The primer extension primers were designed to chimeric sequences obtained from the transposon display experiments however they do not include retroelement sequence.

Table 2-8 displays the oligonucleotides used in the primer extension reactions. All oligonucleotides were supplied by Sigma Genosys (Australia).

2.4.2 Primer Purification

The primer extension primers were purified by running 50 µg of each primer on a 12% polyacrylamide gel and then UV shadowing was used to visualize the bands. Bands were excised, ground up and placed in 500 µl of water and 50 mM NaCl. They were then placed on a rotor at 4°C overnight. The next day samples were spun for 2 min at 13 000 rpm and the supernatant was removed and added to 2 volumes of 100%

ethanol and 1/10 volumes of 3M sodium acetate. The samples were then placed at -80°C for a minimum of 2 ½ hours before being pelleted at 13 000 rpm for 10 min. The supernatants were discarded and the pellets were then washed with 70% ethanol. The supernatants were again discarded and the pellets were allowed to air dry before being resuspended in 70 µl of DEPC water. Finally the primers were run through a Micro Bio-spin 6 column (BioRad, USA) following the protocol provided.

2.4.3 Primer Extension Radiolabelling

The oligonucleotide primers were labelled as described in Section 2.3.5.

2.4.4 Primer Extension method

Total RNA for the reactions was extracted and quantitated as explained in Sections 2.3.1 and 2.3.2. RNA was extracted from the liver or kidney of either an a/a C57Bl/6 mouse, an A^{vy}/a C57Bl/6 mottled mouse or an A^{vy}/a C57Bl/6 yellow mouse. All mice were obtained from the Garvan Institute Animal Facility. Both male and female mice were used and their ages ranged from 3 weeks to 9 months.

Primer extension reactions were carried out following the protocol outlined by Triezenberg (1992) with the some modifications. 5 x 10^5 cpm of end-labelled oligonucleotide was hybridized to 50 µg of total RNA in 10 x hybridisation buffer and this was incubated for 90 min at 65°C before being allowed to cool to room temperature. 30 µl of primer extension buffer was then added to each sample along with 1 µl of SuperScript II RNase H^- RT (Invitrogen, USA) and the samples were incubated at 42°C for 1 hr. The reaction was stopped by adding 100 µl of boiled RNase reaction buffer and 4 µl of Rnase Dnase free and incubating the samples at 37°C for 15 min.

The samples were cleaned up using a standard phenol/chloroform protocol as described in Current Protocols in Molecular Biology Section 2.1.1 (Budelier and Schozz 1992).

Primer Extension Oligonucleotide		Sequence
IAP Chimeras Antisense		
Cpne8PE	(40mer)	5' GTATTACCTCGTCAGAGCCCTCCACTCAGCCCTGCTGTAG 3'
NipaPE3	(40mer)	5' GTAAGGTAGGAACTTTAATGCAAGTTTGAGTGTTGGGAGC 3'
Sec8PE	(39mer)	5' GTTCTTATCGAGACCATATGCTCTTTCCTATCTACTCAC 3'
LINE Chimera Sense		
Plxdc2PE	(39mer)	5' GCTTGACTTTGGTGAAAATGGAGAATAGTTCCCCTGTAC 3'
LINE Chimeras Antisense		
Cugbp2PE	(39mer)	5' CTTAGAATTAAGACATTTGGAACTTAGGAAATTAGTCTG 3'
Unc13aPE	(40mer)	5' CCAAGGCAACTCATAAAAGAAAGAGTTTAGGCCCGGCTGG 3'
MuLV Chimera Sense		
Anxa6PE	(39mer)	5' CCTGTGCTGCCTCCGGGAAGAACTGGCGGGGGTCGGTGG 3'
SINE Chimeras Sense		
Cklfsf8PE	(38mer)	5' CTAACTACTCAAGAGTAAAATAATCCAGAGAGAGATGC 3'
Cops7bPE	(40mer)	5' CAGATTTCAAATGTCCTCAACCTCGTAACCATAATTCTGG 3'
Frap1PE	(39mer)	5' CAGGGCCTACATGCTCTACGTCACAGGCAATGCAGACAG 3'
Kif5aPE	(40mer)	5' GTTTCAAAACACCCAAAAGCAGAATTCATTTATCTTTTGG 3'
Nr6a1PE	(40mer)	5' CAAAAGCATCCTTAAGTTTGATAGATAGAAACTGATTCCC 3'
Pafah1b1PE	(39mer)	5' GTGGACACACACACAAATTAACATAGACTCTGACATAAG 3'
SINE Chimeras Antisense		
Abcd4PE	(38mer)	5' GGCCAGGGTGAGAAAACTATCGGACTCAAGAACTGGTG 3'
Agpat2PE	(39mer)	5' GTACAGCCTGGATCAACAGTGCTCTCAAGAACGCTACCC 3'
AmfrPE	(40mer)	5' GCAGTTTGTGATGCTGTGCTCCTTTGAAACTGACATCCAG 3'
Arih2PE	(40mer)	5' GCAGGAAAGTCTGTAAGAAGCCAACCTGGGCCGGGCATGG 3'
Atad3PE	(40mer)	5' CAGGTTTACCTGAGGAGGCATAAGAGGCCCTTCTGAGCTG 3'
Car5aPE	(39mer)	5' GCACTTTGTGGGCTGAGCTGTTCCTCAGCTCTGCTCGGG 3'
Cdk7PE2	(40mer)	5' GTTGCTTTACTTACTTAGTTAGGTAATCAAGACAGCGTAG 3'
Es1PE	(41mer)	5' CTATGGAATGTCTTATTCACTAAAATGTTACAAGAATGTTC 3'
GnsPE	(40mer)	5' CAGGCCAGAGAGCCCTTGTCTAAGAAACCGAGCTGAGAAG 3'
Ifngr2PE	(40mer)	5' CTTTGCCTGTTTTCCGTAAAGTATTGACTTGTGAAAAGTC 3'
Ncoa1PE	(41mer)	5' CATTATTGCAGAACACTTTTGTGTGTTAAGAACGAAGCATC 3'
Pcmt1PE	(40mer)	5' GGGAGGTCTTATGGAGTAATACTGTACTTGAAGAAATGTG 3'
Pik3PE	(40mer)	5' GAGGCTTGGAGGTCAAATGGTTCGGGCAGTGAAGGGGAAG 3'
Rrn3PE	(40mer)	5' GACGATGTCTCGTTCCGACCGACAAGTCACATTTTTGTCG 3'
Siat4cPE	(39mer)	5' GCACTTGCTACGTAAGCTTGAGGGCTAGCTGGGCAGTGG 3'
Smp1PE	(38mer)	5' CTGCCAGCTTCTATCATGGTCTAAAGAGTTGGTGATGC 3'
VL30 Chimera Sense		
Arhgef12PE	(40mer)	5' GTTTAAAACAAAAGGCCAATAGAAAAGAAAGTGAAAGACC 3'
Cellular Control		
Beta actinPE2	(38mer)	5' GGTGGCGGGTGTGGACCGGCAACGAAGGAGCTGCAAAG 3'

Table 2-8 Oligonucleotides used in Primer Extension reactions

Primer extension products were resuspended in 10 µl of water and 10 µl of display loading dye, before being run at 1800 V on a 5% polyacrylamide/7 M urea sequencing gel until the bromophenol blue dye was 2/3 of the way down the gel. Bands were visualized by autoradiography overnight at -80°C

2.5 RT-PCR

2.5.1 Oligonucleotides

The oligonucleotide primers for the RT-PCR experiments were designed to chimeric sequences obtained from the transposon display experiments. Table 2-9 shows the oligonucleotides used in the RT-PCR experiments. All oligonucleotides were supplied by Sigma Genosys (Australia).

RT PCR Oligonucleotides		Sequence
EST Database Candidates		
IL25PCRfwd	(19mer)	5' GGTGGCTCTGTGGGGTCAG 3'
IL25RT1	(29mer)	5' CCATCTCTCTAACCCTTTGTTTTATTTAC 3'
JupPCRfwd	(21mer)	5' CACAACCTGCTGCTCTATCAG 3'
JupRT1	(24mer)	5' CCCGACCCCTTCCCTCTTTTCATC 3'
Snrp1cPCRfwd	(22mer)	5' CCATTCACAGCTGCGACTGAAC 3'
Snrp1cPCRfwd2	(19mer)	5' TGGGCTGCCGACTACGGAG 3'
Snrp1cRT1	(26mer)	5' CAATCAGGCTCTGGGCCTGCTCTTCC 3'
IAP Chimeras		
Cpne8PCRfwd	(20mer)	5' CTTCCCACACCCTACAGCAG 3'
Cpne8PCRfwd2	(27mer)	5′ ACACCCTACAGCAGGGCTG 3′
Cpne8RT1	(26mer)	5' TGCTGAGGTTCAATAGACCATTCCGC 3'
MuLV Chimeras		
Anxa6PCRfwd	(19mer)	5' GACCCCCGCCAGTTCTTCC 3'
Anxa6RT1	(24mer)	5' GCCAGGTCTCCCGAGATCTCTGAC 3'
SINE Chimeras		
Car5aPCRfwd	(19mer)	5' CCACGTGGCACGACGAGCC 3'
Car5aRT1	(24mer)	5' GTTCCTCAGCTCTGCTCGGGTTTC 3'
Cdk7PCRfwd	(21mer)	5' CACCCAGCTTTGTTCTCTGAG 3'
Cdk7RT1	(24mer)	5' GGCTCCCAAATGATTTGGCCAGGC 3'
Nr6a1PCRfwd	(19mer)	5' CCCTGTCTCTGGCTTGGGG 3'
Nr6a1RT1	(25mer)	5' GATAAGCCCAACTCCTTCAGGGAGC 3'
Pik3PCRfwd	(19mer)	5' CCACTGCCGGCTTCTTCCC 3'
Pik3RT2	(25mer)	5' GTGGTCAGAGCATCAGGAGACTGAG 3'

Table 2-9 Oligonucleotides used in RT-PCR experiments

2.5.2 RT-PCR method

Total RNA was extracted and quantitated as explained in Section 2.3.1 and 2.3.2. Oligotex mRNA Kits from QIAGEN (Germany) were then used to extract mRNA from the total RNA. RNA used in the RT-PCR experiments was extracted from the liver and kidney of Balb/c, C57BL6J, DBA, Fvb, Q/s and 129SvJ mice. All mice were obtained from the Garvan Institute Animal Facility. Both male and female mice were used and their ages ranged from 3 weeks to 9 months.

The RT-PCR reactions were carried out using the SuperScript One-Step RT-PCR with Platinum *Taq* kit from Invitrogen (USA). The provided protocol was followed with 50 ng of template mRNA being added to each reaction and 2 U of Taq DNA Polymerase (Roche Applied Science, Germany) being added to the control tubes. The PCR amplification step was carried out using the suggested temperatures with an annealing temperature of 59°C and 35 cycles being performed.

10 µl of each product was run at 110 V for approximately 30 min on a 1% agarose TAE gel, with 0.3 µg/ml ethidium bromide, before bands were visualized and recorded using the Quantity 1 – 4.1.0 software on the GelDoc 2000 (BioRad, USA).

For the Cpne8 and Snrp1c samples an additional nested PCR step was necessary to produce a specific band. In each case a 1/100 dilution was made of the first PCR product and 5 µl of this was added with 1 µl of each of the forward and reverse primers to a PCR mix consisting of 10 mM Tris-HCl, 50 mM KCl pH 8.3 buffer, 2.5 mM $MgCl_2$, 0.05 mM dNTPs and 1 U of Taq polymerase in a 50 µl final volume. 10 µl of each product was run on a gel to analyse results as described above.

CHAPTER 3 IN SILICO ANALYSIS

3.1 Introduction

The aim of this chapter was to provide evidence that retroelements have intact regulatory regions which are capable of becoming active. While previous studies had reported transcriptional expression of a variety of retroelement families in both the human and mouse, no large scale study had been conducted to see how many families were capable of transcription. An additional aim was to identify whether retroelement transcription in humans was restricted to diseased tissue, as was commonly believed, or whether transcription is in fact equally as common in normal healthy tissue. Finally we tried to observe patterns of retroelement expression, by conducting a preliminary study aimed at identifying chromosomes where retroelement transcription was abundant.

Unlike single copy genes, retroelements are difficult to study; this is due not only to their large numbers but also to sequence conservation between family members. For example the mouse L1Tf LINE family is present in about 3000 full length copies per diploid genome with members sharing greater than 99% sequence similarity (Ostertag and Kazazian 2001). While not all retroelement families are present in such large numbers, and most do not display such a high level of sequence similarity, it is still difficult using standard techniques to distinguish between individual family members. In order to discover supporting evidence for our hypothesis that retroelements may be capable of exerting transcriptional interference on surrounding genes, we had to overcome this problem.

With the rapid expansion of the expressed sequence tag (EST) databases during 1999-2002 (Marra *et al.* 1999; Dias Neto *et al.* 2000; Okazaki *et al.* 2002), papers were being published which utilised this new data via computer BLAST methods. EST searches produced sequences of new proteins and genes (Pandey and Lewitter 1999; Schultz *et al.* 2000), as well as sequences for new retroelements (Bromham *et al.* 2001).

In addition the expansion of Repbase Update, a database containing representative examples of repetitive elements (Jurka 2000), and the formation of the repeat BLAST program Repeatmasker, meant identifying retroelements in sequences would be much easier than was previously possible. These findings led us to consider conducting an *in silico* analysis to address our hypothesis, as it seemed likely that a computer based method would allow us to study individual retroelements.

For retroelements to be acting as controlling elements, and therefore exerting transcriptional interference, they would need intact and active promoters. Retroelements which have active promoters would be expected to produce transcripts and in turn these transcripts should be present in the EST database (dbEST). We decided that by conducting an *in silico* analysis that mined the dbEST (which would provide a large pool of examples versus just examining a single library), we might find evidence that there are active retroelement promoters in the human and mouse genomes.

Shortly after we started the *in silico* analysis, the human genome sequencing effort was completed (Venter *et al.* 2001) and we realised this would be a resource which would allow us to find chromosomal placements of individual retroelements. When the mouse genome was completed in 2002 (Waterston *et al.* 2002) we decided to broaden our analysis and examine expression of the mouse retroelement families.

Since retroelement expression has hypothetically been linked to diseases such as cancer (Bratthauer *et al.* 1994; Ristevski *et al.* 1999), schizophrenia (Yolken *et al.* 2000) and diabetes (Krieg *et al.* 1992) we were interested in analysing the recovered transcripts to see whether expression of retroelements was restricted to diseased tissue, or if their expression was equally common in normal tissue.

The aim of this chapter is to,

- try to determine how many human retroelements have intact and active promoters via a computer based analysis which utilises both dbEST and the human genome database.
- determine if retroelement expression is more common in diseased versus normal tissue.
- look at the chromosomal location of the active retroelements to observe if there is a pattern of expression.
- apply the first aim to mouse retroelements to see if their activation differs from human retroelements.

3.2 Results

In order to obtain an overall view of retroelement activity in the human genome, all of the retroelement families described in the 8.1.5 version of RepBase (the online repetitive sequence library database) were included in the human retroelement analysis. The mouse retroelement analysis was conducted in less detail with only the main murine retroelement families being examined (refer to Section 3.2.5).

Three different types of probes were used for the human and mouse retroelement analysis, depending on whether the retroelement could be broadly categorised as a LINE, SINE or ERV retroelement. Probes for ERV retroelement BLASTs were sequences taken from the transcription start site or R region. In most cases the 5' LTR R sequence was used for the probe as this sequence is generally considered to be better conserved than the 3' LTR R sequence, due to the way the ERV elements replicate. The promoter sequence (for humans) or monomer sequence (for mice) was used as a probe for the LINE elements while the whole SINE element was used as a probe. Refer to Figure 3-1 and Figure 3-2 for a schematic of the probes used. Accession numbers for dbEST probes are provided in Chapter 2.

The resulting ESTs were kept for further analysis if they had initiated or terminated in the R region of ERV retroelements, at either end of the SINE elements, or in the promoter region of the LINE elements. ESTs which did not meet these criteria were discarded. The ESTs were then categorised according to similarity to the probe sequence, sense or antisense orientation and the library from which they were derived. This strategy entailed a great deal of labour, since each EST recovered had to be examined and assessed. This should be kept in mind when viewing the final products of the analysis. Since an EST has a 5' and 3' end (determined by the position of the poly A tail), a search was then conducted to determine if both ends of the EST

had been sequenced. Finally the ESTs were used as a probe to BLAST the human or mouse genome in order to determine the chromosomal location.

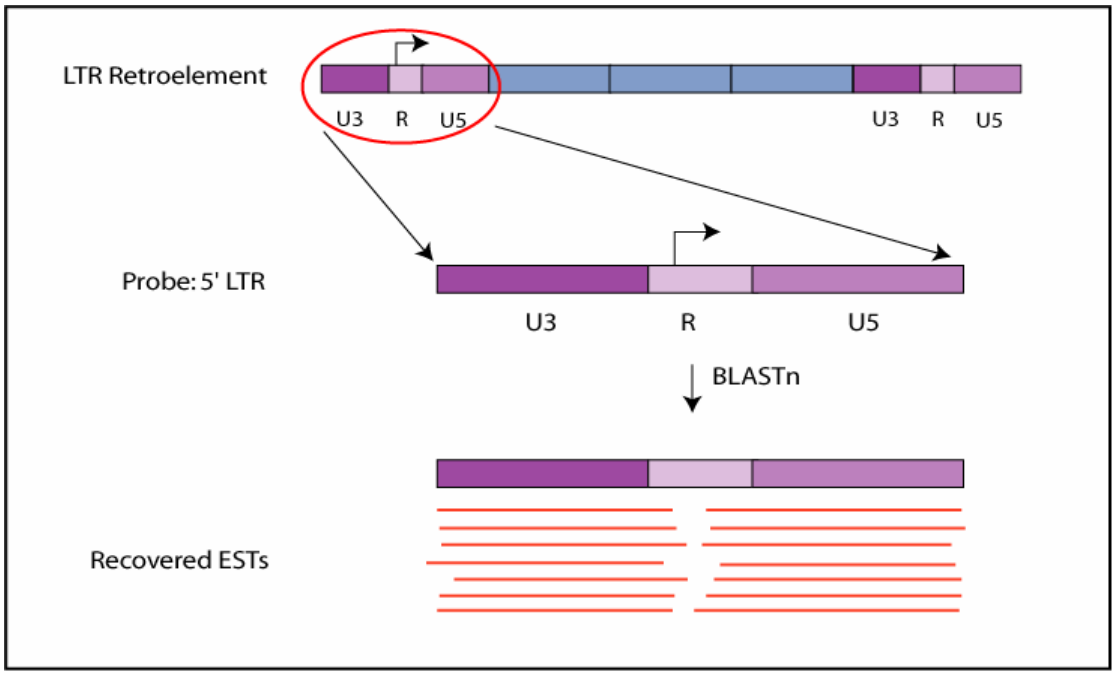

Figure 3-1 Strategy for dbEST search for ERV retroelements. Sequence from the LTR region, including the start site of transcription (R), was used as the BLASTn probe. The NCBI BLASTn program was used with a limit of 1000 returned sequence descriptions. Recovered ESTs were kept for further analysis if the 5' or 3' end began in the R region.

A detailed description of the BLAST method, including the Genbank and Repbase accession numbers for probes, may be found in Chapter 2.

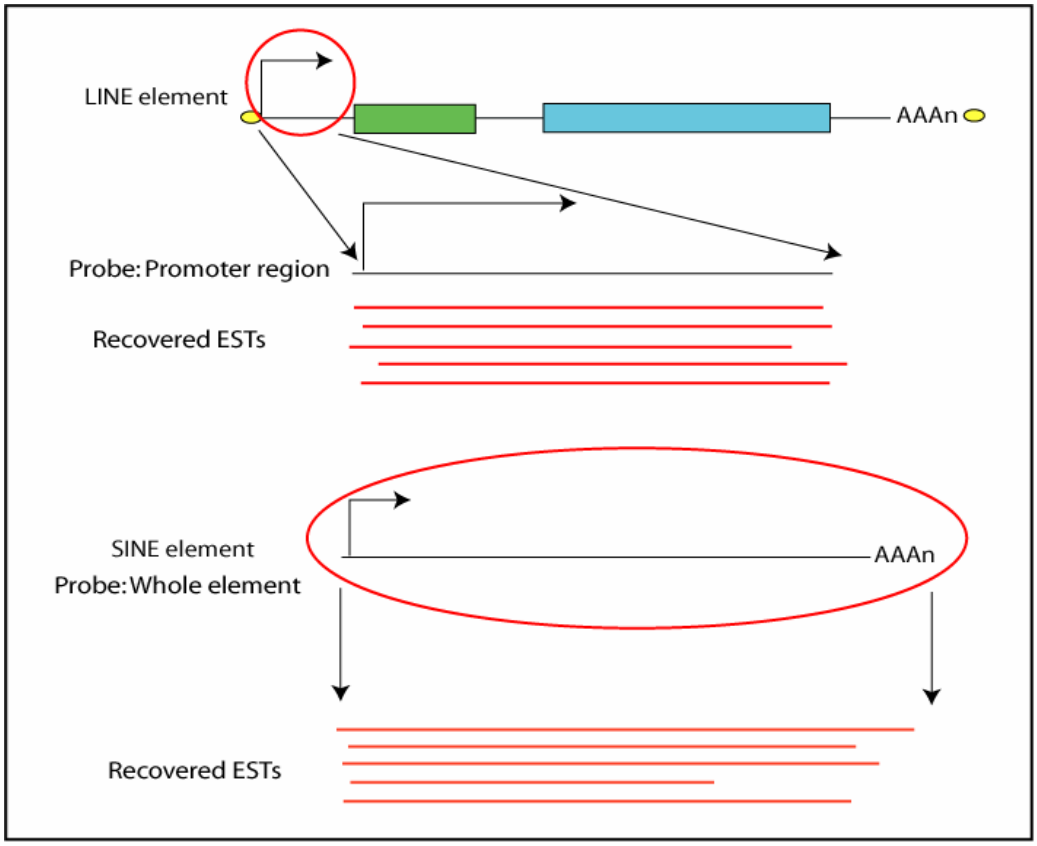

Figure 3-2 Strategy for dbEST search for LINE and SINE retroelements. Probes for the BLASTn program were sequence from the promoter region for the LINE elements or the whole sequence for the SINE elements. The NCBI BLASTn program was used with a limit of 1000 returned sequence descriptions. Recovered ESTs were kept for further analysis if the 5' or 3' end began in the promoter region of the LINE element or at either end of the SINE element sequence.

3.2.1 Data Analysis

The large amount of data we obtained from the *in silico* analysis could be examined in numerous ways. We chose to focus specifically on the number and orientation of retroelement ESTs in the database for both humans and mice, the number of active retroelements detected in the human genome, the chromosomal location of the

active retroelement promoters and the types of libraries (tissues) from which the human ESTs were derived.

3.2.2 Human Retroelement Analysis: Results

For the majority of the retroelement families examined, a large number of ESTs were found which fit the specified inclusion criteria. The ERV family of elements had the largest number of ESTs included in the analysis: all except two ERV families, ERV-1 and HTLV-1, had ESTs recovered which were included. The lack of ESTs recovered for the ERV-1 and HTLV-1 families does not necessarily mean the families are not active. The recovered ESTs may not have fit the inclusion criteria and may therefore have been excluded even though they may have been legitimate transcripts initiated/terminated within the retroelement. The SINE elements had the second largest number of transcripts recovered with the LINE elements returning the lowest number of transcripts. Refer to Table 3-1 for the human retroelement analysis results.

A subset of ERV families such as HERV-L and sub families of HERV-K were not included in the analysis, as it was unclear where the start site of transcription resided and therefore they did not meet the inclusion criteria. Likewise many of the less well studied retroelement families such as those in the MaLR group, were also not included. These families are considered to be quite active, and therefore overall numbers would be higher if they had been included in the analysis.

RETROTRANSPOSON FAMILY	GENBANK ACCESSION NO./REPBASE ID FOR PROMOTER PROBE	NO. OF TRANSCRIPTS			NO. OF ACTIVE LTRS/ELEMENTS
		Sense	Antisense	Unknown	
LINE	Repbase release 8.1.5 L1 consensus (1-1030 bp)	269	226	262	409
SINE	Repbase release 8.1.5 Alu consensus (1-290 bp), AluJo (1-283 bp), AluSc (1-280 bp), AluSp (1-284 bp), AluSx (1-283 bp), AluY (1-281 bp), AluYb8 (1-288 bp) and AluYd2 (1-270 bp)	668	163	341	N/D
ERV Families					
ERV-1	K02919 (1-480 bp)	0	0	0	0
ERV-9	AF064191 (421-982 bp)	187	16	65	113
HERV-E	M10976 (251-650 bp)	119	0	5	23
HERV-F	Kjellman *et al.* 1999 5' LTR (301-536 bp)	28	0	3	1
HERV-H	AF108838 (301-700 bp)	216	4	90	93
HERV-I	Repbase release 8.1.5 LTR10A (301-609 bp)	37	0	2	8
HERV-K	Repbase release 8.1.5 LTR3 (1-432 bp)	102	9	36	33
HERV-P	Repbase release 8.1.5 LTR8 (401-691 bp)	52	4	8	25
HERV-R	K02016 (361-618 bp)	47	0	1	2
HERV-W (MSRV)	AF127226 (201-600 bp)	48	0	6	15
HTLV-1	X16660 (1-684 bp)	0	0	0	0
RRHERV-I	Repbase release 8.1.5 LTR15 (241-493 bp)	28	1	6	17
S71	Repbase release 8.1.5 LTR6B (301-558 bp)	81	1	7	18
	TOTAL	1882	424	832	757

Table 3-1 Results for the human *in silico* analysis. Transcripts are shown categorised by family type, orientation and by the number of LTRs/elements found to be active. A number of ESTs were not annotated for transcript orientation so these sequences have been included in the unknown category. Due to the sequence similarity of the SINE elements, their chromosomal location could not be determined so the number of active elements could not be determined. N/D – not determined.

Most of the recovered transcripts were sense direction transcripts (1882) with a much lower number of transcripts found to be antisense (424). **The number of antisense transcripts recovered is substantial enough however to suggest antisense transcription is actually occurring from some promoters, and the transcripts are not just examples of truncated transcripts that are initiated elsewhere.** This is supported by the previously described A^{vy} IAP and the multiple L1 examples described by Speek (2001) and Nigumann *et al.* (2002), which are all examples of antisense transcription from retroelement promoter regions.

The number of transcripts which could not be assigned an orientation, due to incomplete information was quite high (832). They could not be used for some of the further analysis, but were included in the study as they met inclusion criteria and were therefore likely to be authentic transcripts.

As may be expected, the number of active LTRs/elements is lower in number than the total number of transcripts recovered for a family. This is because in some cases a transcript was present in more than one copy – sometimes multiple copies of one transcript were found in one library while in other cases the same transcript was found in different libraries.

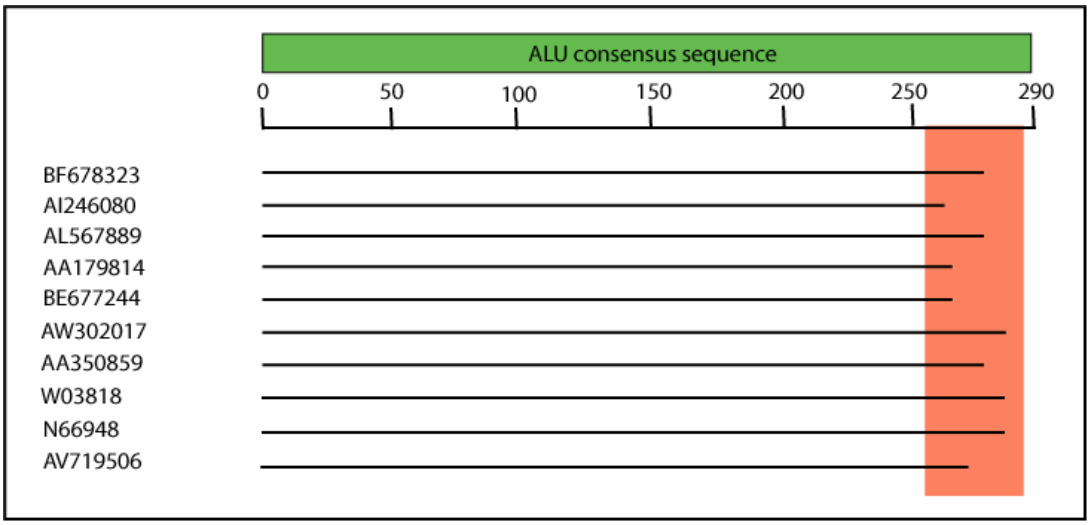

Figure 3-3 Graphical representation of a subset of Alu antisense ESTs. The Repeatmasker Alu consensus sequence is represented as a green bar and the antisense ESTs are represented as black lines. The Genbank accession numbers are alongside the ESTs and the likely region of the antisense promoter is shown shaded orange. Only a small subset of antisense sequences which were returned from the analysis are shown.

While it is widely recognised that ERV retroelements have two possible promoter regions (one in the 5' and one in the 3' LTR), and there is evidence that L1 elements have a sense and an antisense promoter (Speek 2001; Nigumann *et al.* 2002),

an antisense promoter in the 3' end of SINE elements has not previously been described. Our analysis recovered transcripts which appear to be initiated in the very 3' end of SINE elements **suggesting an antisense promoter may reside in that region**. Refer to Figure 3-3 for an example of antisense SINE transcripts.

3.2.3 Retroelement Transcript Tissue Distribution

Studies such as Nakagawa *et al.* (1997) and Yolken *et al* (2000) have suggested retroelement expression is more common in diseased tissue than normal healthy tissue, however the results we obtained suggest this may not be the case. The recovered ESTs were categorised as being derived from either cancer, normal or placental libraries (kept in a separate category as they are neither maternal nor fetal tissue) and our study found that the number of ESTs derived from malignant and normal libraries were very similar. There were lower numbers of placental ESTs recovered, but this is likely to be due to the smaller number of placental libraries present in the dbEST at the time (refer to Figure 3-4).

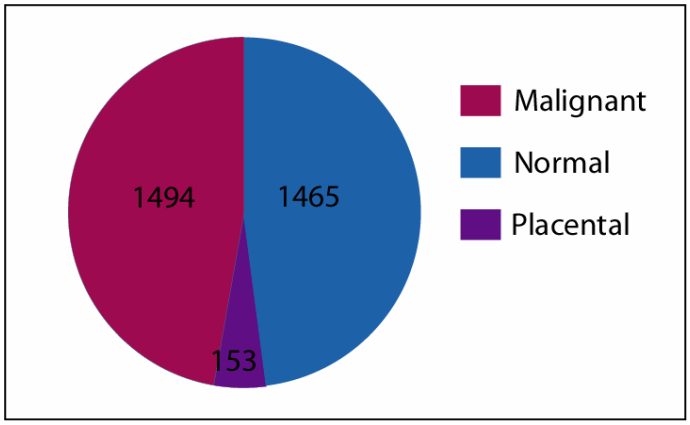

Figure 3-4 Human retroelement BLAST results categorised by EST tissue type. Retroelement ESTs were placed into three main groups according to the type of tissue the EST had been derived from. There were similar numbers of ESTs derived from malignant and normal libraries while a substantially lower number of ESTs were derived from placental libraries. For more information on which libraries the transcripts were derived from – refer to the Appendix.

It should be stressed that the results are raw data so they have not been adjusted in any way – a more complete analysis would need to be done to confirm tissue patterns of expression of retroelements.

3.2.4 Distribution of Human Retroelement Transcripts in the Genome

In order to examine if the active human retroelements were located in similar regions, we assessed the distribution of unique transcripts in the genome (refer to Figure 3-5. This preliminary analysis included transcripts which were initiated or terminated in a retroelement regulatory region.

Generally there appears to be a slight downward trend in the number of active LTRs/elements as the chromosomes decrease in size, with the exception of chromosome 19 which has a large number compared to its size. Apart from this there does not appear to be an obvious pattern. It is interesting to note that chromosome 19 has the highest gene-based marker assignments of any human autosome (Croft *et al.* 1999), which may help explain the higher than expected number of active retroelements on chromosome 19. If retroelements are being activated due to chromatin and methylation changes induced by nearby genes, then it would be expected that a chromosome with a high density of genes would also have a large number of retroelements transcriptionally active.

Most of the retroelement families examined displayed varying amounts of transcript expression from locations on approximately half of the chromosomes. The ERV-9 family was the only family which produced transcripts from positions on all of the chromosomes, while the HERV-H and LINE families were expressed from all except chromosome 1 and 4 respectively. The HERV-F family was only expressed from a single location on chromosome 7 while HERV-R was only found to be active on chromosome 7 and 22.

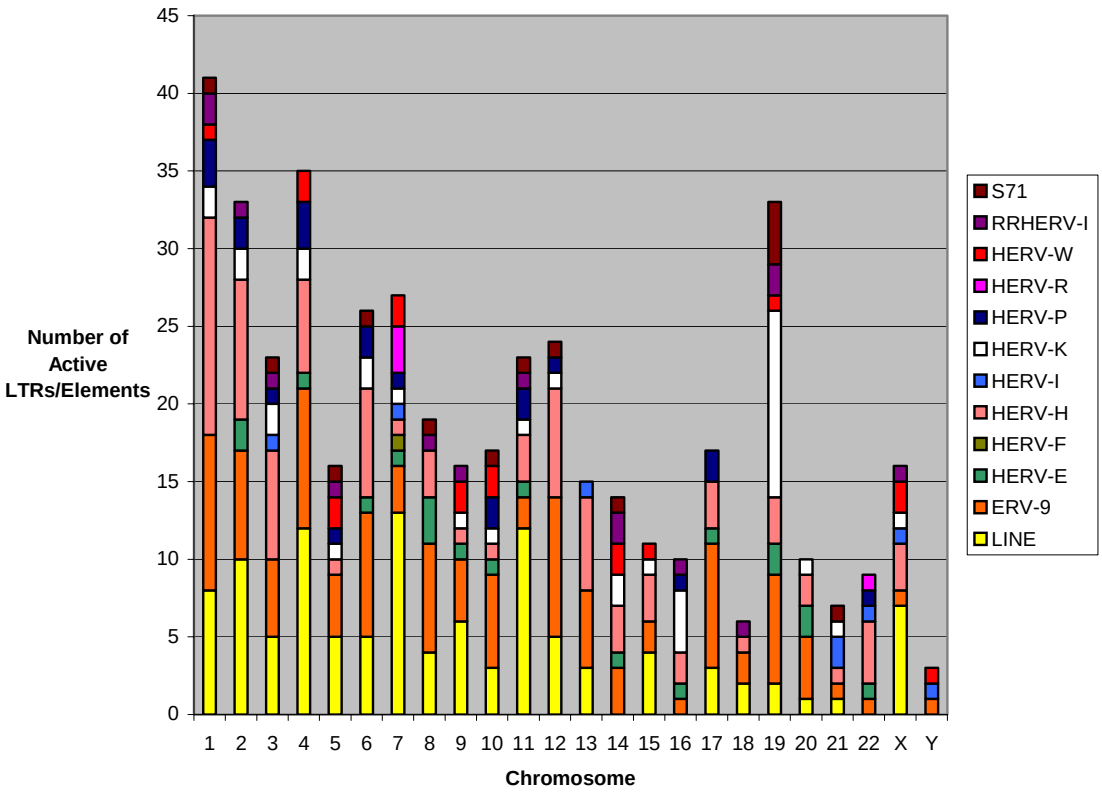

Figure 3-5 Chromosomal distribution of unique human retroelement transcripts. All the retroelement families examined, which produced transcripts, were included in the graph below except for the SINE family which did not have chromosomal locations assigned to transcripts due the high sequence similarity. The ERV-1 and HTLV-1 families did not have any transcripts recovered. The HERV-K subfamily examined was LTR57. The transcripts were located in a range of different libraries – for detailed information refer to the Appendix.

Most of the retroelement families examined displayed varying amounts of transcript expression from locations on approximately half of the chromosomes. The ERV-9 family was the only family which produced transcripts from positions on all of the chromosomes, while the HERV-H and LINE families were expressed from all except chromosome 1 and 4 respectively. The HERV-F family was only expressed from a single location on chromosome 7 while HERV-R was only found to be active on chromosome 7 and 22.

Chromosome 1 had the most unique retroelement transcripts at 41, followed by chromosome 2, 4 and 19 with 33, 35 and 33 unique transcripts respectively. Chromosome Y had the least number of unique transcripts with only 3 being located. The average number of unique transcripts per chromosome was 18.

The SINE elements were not included in the genome analysis as the sequence similarity is very high for the family, and there are 1.5 million copies in the genome which made locating a single element difficult. Since the study was conducted, advances have been made with BLAST programs which may allow for SINE elements to be located, but at the time of this analysis it was not possible. Unique transcript numbers would have been much higher if these elements had been included.

As mentioned in Section 3.2.2 other active families were also excluded due to the difficulty in recognising if a transcript had been initiated/terminated within a promoter/stop region so the numbers we obtained are likely to be an underestimation.

Again it should be stressed that the results are raw data so they have not been adjusted in any way – a more complete analysis would need to be done before strong conclusions are drawn.

3.2.5 Mouse Retroelement Analysis Results

The mouse retroelements were not examined in as much detail as the human retroelements due to the fact that less information on the mouse retroelement families was available at the time the study was conducted. Therefore only the main retroelement families – LINE, SINE, ETn, GLN, Merv-L, MMTV, MuERV-C, MuLV, MuRRS, MuRVY and VL30 were included in the analysis.

As was seen with the human analysis, the ERV family had the largest number of ESTs included in the analysis with the SINE and LINE families also producing large numbers. ESTs were not recovered for three of the ERV families examined – MuERV-

C, MuRRS and MuRVY, while the IAP family was not included in the analysis due to the large variation in the start site of transcription (Kuff and Lueders 1988), which made determination of authentic transcripts difficult.

RETROTRANSPOSON FAMILY	GENBANK ACCESSION NO./REPBASE ID FOR PROMOTER PROBE	NO. OF TRANSCRIPTS			NO. OF 5' TRANSCRIPTS
		Sense	Antisense	Unknown	
LINE	Naas et al.1998 L1Tf (1-374 bp) , M29324 L1A (1092-1837 bp), AC068252 L1Gf (161291-159839 bp),	194	123	12	133
SINE	Repbase release 8.1.5 B1 (1-135 bp), B2 (1-209 bp), B3 (1-226 bp)	429	73	36	112
ERV Families					
ETn	X03063 (1-322 bp)	62	14	1	22
GLN	M14005 (1-618 bp)	135	7	5	70
MeRV-L	Y12713 (250-530 bp)	117	13	2	78
MMTV	AF228550 (900-1480 bp)	310	4	5	230
MuERV-C	AF049340 (5661-6234 bp)	0	0	0	0
MuLV	X05157 (480-720 bp)	111	2	1	70
MuRRS	X02487 (1-509 bp)	0	0	0	0
MuRVY	M27506 (1-627 bp)	0	0	0	0
VL30	M21123 (200-500 bp)	128	15	0	39
	TOTAL	1486	251	62	754

Table 3-2 Results for the mouse *in silico* analysis. Transcripts are shown categorised by family type, orientation and by the total number of 5' transcripts recovered.

The majority of the mouse retroelement ESTs were oriented in a sense direction (1486) with a much smaller number being antisense (251) or of an unknown orientation (62). Close to half of the transcripts recovered were initiated in the promoter region of a retroelement, suggesting that a large number of retroelement promoters are capable of transcription. Refer to Table 3-2.

It was interesting to note that approximately 45% of the transcripts were obtained from embryonic and neonatal libraries (see Appendix). This perhaps is to be expected since 40% of the mouse libraries were made from embryonic or neonatal tissues, at the time the analysis was conducted. This means that there would be a high chance that retroelements would be found in this type of library simply due to the

abundance of embryonic and neonatal libraries. However this finding may also highlight the fact that the retroelements are becoming active when methylation is at its lowest in a cell. Further analysis would be needed to be able to distinguish between the two possibilities.

As was observed in the human analysis, the mouse SINE elements appeared to have an antisense SINE promoter. The region varied for the different mouse SINE families; Figure 3-6 only shows the B2 family which has a well defined region for the likely antisense promoter region.

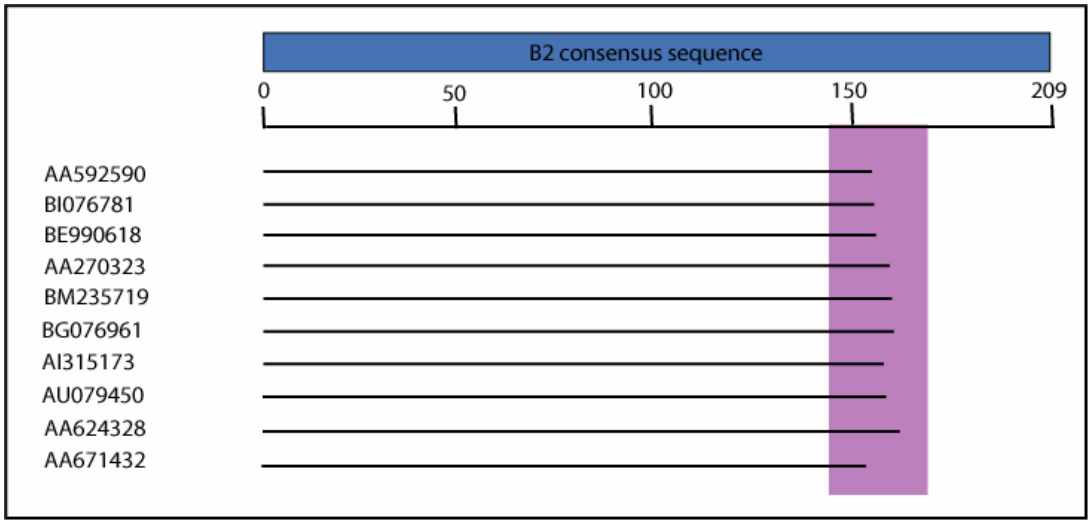

Figure 3-6 Graphical representation of a subset of B2 antisense ESTs. The Repeatmasker B2 consensus sequence is represented as a blue bar and a small subset of antisense ESTs recovered from the analysis are represented as black lines. The Genbank accession numbers are alongside the ESTs and the likely region of the antisense promoter is shown shaded purple.

3.2.6 Chimeric Transcripts

During the BLAST search of the human and mouse genomes, it became apparent that a small number of transcripts appeared to initiate/terminate within a retroelement and be joined to cellular gene sequence. In some examples, such as the

Snrp1c example in Figure 3-7, the cellular sequence was spliced to the retroelement sequence. In the case of Snrp1c, the Merv-L promoter is 1.9 kb upstream of the gene suggesting the chimera may be an example of legitimate splicing, and not just a case of chance splicing due to elements being in close proximity. The Jup chimera is an example of a B1 element appearing to act as an alternate stop site terminating the transcript after exon 4, while the last two examples from the human analysis, SERPINF1 and PLAUR, are both Alu elements appearing to initiate transcription from within an intron.

These are obvious examples of transcriptional interference where the retroelement is interfering with a nearby gene - either by transcribing into cellular sequence, being spliced to cellular sequence, or terminating the cellular transcript.

The chimeras described above are just a few of the examples which were recovered from the analysis; there are likely to be many more examples within the dbEST but the chimeras were not exhaustively analysed due to the time required to achieve such an aim.

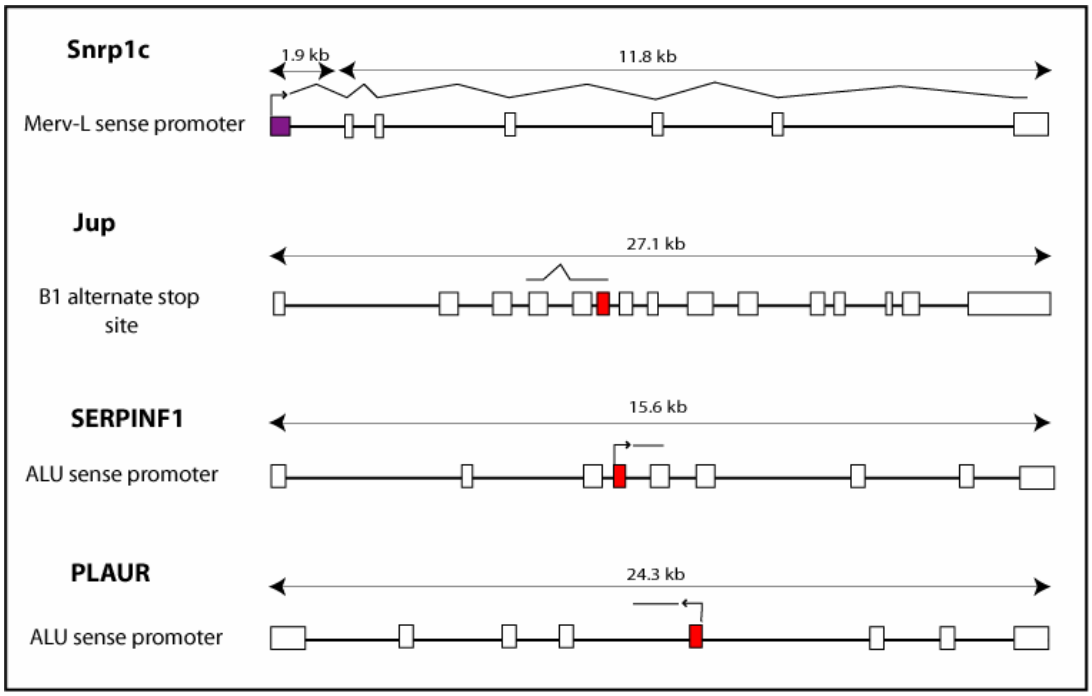

Figure 3-7 An example of chimeric transcripts recovered from the mouse and human genome searches. Four types of chimeric transcripts are shown below – the first two are from the mouse analysis and the second two are from the human analysis. The first example shows a retroelement 5' to the gene acting as an alternate start site of transcription for the gene. This chimera was found in the NIA Mouse 15K cDNA Clone Set made from pooled normal embryos. The second example shows a retroelement acting as an alternate stop site for the cellular gene transcript. This chimera was found in the Stratagene mouse skin (#937313) library made from skin from 11 week females. The third example shows a retroelement transcribing out of intron into a nearby exon and it was found in the NIH_MGC_118 blood leukocyte library. The final example shows a retroelement transcribing in an antisense direction to the cellular transcript and it which was found in the 313 (synonym: hlcc2) uncharacterised adult tissue library.

3.3 Discussion

Both the mouse and the human retroelement analyses returned a large number of transcripts which were initiated or terminated within the regulatory region of a retroelement. On the whole a larger number of transcripts were recovered in the human analysis, but this is likely to be due to the fact the mouse study was not as thorough. It is widely believed that retroelements are always methylated and maintained in a silent state; my finding that some regulatory regions are active suggests this view is incorrect.

Results for the human and mouse analyses were quite similar. The largest number of transcripts recovered in the human and mouse analyses were for the ERV family (1209 and 934 transcripts respectively) followed by the SINE elements (1172 and 538 respectively) and finally the LINE elements (757 and 329 respectively). There may appear to be a discrepancy when ratios of recovered transcripts are compared to the copy number for the three families - in the human genome there are approximately 1,500,000 SINE elements, 850,000 LINE elements and 450,000 ERVs and in the mouse genome there are approximately 1,500,000 SINE elements, 660,000 LINE elements and 631,000 ERVs. However the total number of transcripts is likely to deviate from copy number of retroelements, as one very active retroelement could be responsible for a large number of transcripts which have all originated in one promoter. Therefore it is better to compare copy number to the number of active LTRs/elements, which is only possible for the human analysis. In the human analysis the ratio of active retroelements to copy number still shows a discrepancy with the ERV family still providing a larger number of active elements than the LINE family (the SINE elements were not included in this analysis). The sizes of the ERV and LINE elements are similar (6-11 kb and 7 kb respectively) so it is unlikely the discrepancy is due to a transcriptional preference based on size difference. Perhaps the discrepancy is due to the fact that the analysis did

not recover all the active LINE elements or perhaps the inclusion criteria resulted in a number of legitimate LINE transcripts being excluded. Alternatively the LINE elements are quite different in structure to the ERV elements and may therefore be silenced in a more efficient manner than the ERVs. Yet another reason is that perhaps the ERV elements are being regulated via binding sites in the LTRs that are not present in the promoters of LINE elements, and this is causing the ERV elements to be transcribed at a higher rate than the LINE elements.

It has also been shown that LINE elements integrate into AT rich sequence while ERV elements integrate uniformly throughout different GC sequence content, avoiding only higher GC% regions (Venter *et al.* 2001). Therefore it is likely that the ERV and LINE elements would be located in different regions; if they were to be activated due to local chromatin changes, then different overall numbers would be activated, with ERV elements being activated in higher numbers since they reside with genes in higher GC% content sequence. However the most plausible explanation for the difference would be that since only the L1H family of LINE elements are thought to be active in the genome, and there are only approximately 1000 copies, there are probably more LTR families which have remained active, and therefore they can produce more transcripts. A more thorough analysis would be needed to answer this question.

Antisense retroelement transcripts were recovered in both the human and mouse studies at a level which was approximately 1/5 that of the sense transcripts. This seems quite a high proportion to just be attributed to random background transcription. Papers such as Yelin *et al.* (2003), Chen *et al.* (2005), Cheng *et al.* (2005) and Katayama *et al.* (2005) suggest antisense transcription in the human genome is commonplace with estimates of well over 20%, perhaps even approaching 40%, of human genes forming sense/antisense (SA) pairs (Yelin *et al.* 2003; Chen *et al.* 2005;

Cheng *et al.* 2005; Katayama *et al.* 2005). Chen *et al.* (2005) goes so far as to claim that many of the SA pairs are co-expressed and inversely expressed more frequently than is expected by chance, and that combined with their evolutionary conservation, this suggests they play a role in gene regulation. The type of regulation which is commonly believed to be occurring is via mechanisms such as RNA interference, where the sense transcript is degraded, as well as by gene silencing via short interfering RNA (siRNA), methylation and chromatin changes. Gene silencing via siRNA involves the processing of repeat RNA transcripts into siRNA of 21-23 nucleotides by Dicer. These fragments then target homologous RNA sequences, which leads to transcriptional silencing. Silencing has been proposed to occur via the siRNA targeting spurious nascent RNAs traversing the promoter region or LTR, which induces DNA methylation and transcriptional silencing leading to heterochromatin formation (Lippman and Martienssen 2004; Schramke and Allshire 2004). This type of silencing has been observed in plants such as Arabidopsis, as well as organisms such as yeast and Drosophila, but while it is likely to be occurring in mammals, it has yet to be convincingly shown to be part of mammalian biology.

Another possibility for antisense transcript function, presented by Kapranov *et al.* (2005) in relation to noncoding transcripts, is that they may function to increase accessibility of regulatory machinery to various coding regions of the chromosome, thereby facilitating transcription. While retroelement transcripts are largely ignored, it is possible these regulatory affects might also apply to antisense retroelement transcript expression, therefore supporting the idea that the antisense transcripts have biological function.

The ratio of sense and antisense transcripts was shown to differ between LINEs SINEs and ERVs (see Table 3-2) with the LINE and elements having a much larger

number of antisense transcripts than sense transcripts. One explanation is that the antisense promoter in the LINE may be more efficient than the one in the SINE and ERV elements. The antisense SINE promoter is yet to be characterised so until then the efficiency of the promoter can not be investigated, while the antisense promoter in the ERV element, while being generally very similar in sequence to the sense promoter, is often degenerate in some degree which could explain the difference in strength. Alternatively the LINE antisense promoter may escape remethylation more easily than the SINE and ERV antisense promoters, due to the differences in sequence between promoters of the three families. Ultimately the cause of the difference in ratios of sense and antisense transcripts of the three families is yet to be defined.

An interesting finding from our *in silico* analysis was that **SINE elements in both the human and mouse appear to have promoters in their 3' ends**. This is perhaps unsurprising considering that ERV and LINE elements have antisense promoters (Speek 2001), however no previous evidence has suggested the existence of a SINE antisense promoter. This finding is of interest, as it suggests that there may be a large number of potentially active promoters in the human and mouse genomes which have previously been overlooked. In terms of transcriptional interference, there are a maximum of 1,500,000 previously unexamined promoters which may be interfering with surrounding genes. Of course, the number would not be this high as most would be silent and many may be incapable of transcription due to deletions or truncations in the promoter region, but it still opens the way for more candidates for transcriptional interference. Further work involving placing the 3' end of the SINE element in a reporter construct, mutating the end and mapping where the promoter lies, need to be conducted before we can conclusively say that SINE elements have an active 3' promoter.

Transcriptional interference has been a focus of our lab for many years, so the transcripts which were of most interest to us were the transcripts initiated in the promoter region of retroelements. We define transcriptional interference as being the influence of one transcriptional unit on another unit linked in *cis.* Therefore if the retroelement is acting as an active transcription unit and is capable of producing transcripts, it may exert transcriptional interference upon a cellular gene. The chimeric transcripts we recovered are likely candidates for this type of interference as they are direct evidence of retroelements interfering with the normal transcription of a gene. The retroelement chimeras may interfere with normal gene function by causing truncated proteins to be translated, as in the case of the $Axin^{fu}$ mice (Rakyan *et al.* 2003), by causing gene silencing to occur via antisense/sense transcript as described above, or by just altering the regulation of the gene by upregulating or downregulating its transcription as seen with A^{vy} mice (Duhl *et al.* 1994). Due to time constraints the chimeras were not extensively studied, but our preliminary search suggests retroelement chimeras may be expressed in low levels throughout the genomes of humans and mice.

While transcripts which were initiated in promoters of retroelements were of most interest, a large number of the transcripts recovered from the analysis terminated in the regulatory regions of retroelements. Some of the transcripts were chimeric, and may be examples of a type of transcriptional interference in which a gene is terminated prematurely, but it is likely that many are examples of intergenic transcription. The function of intergenic transcription is not well understood, but it has previously been described in the β globin locus (Ashe *et al.* 1997); more recently it has been suggested that this type of transcription is commonplace in the human genome (Cheng *et al.* 2005; Kapranov *et al.* 2005). It has been speculated that intergenic transcripts may be capable of developmental regulation (Ashe *et al.* 1997) and therefore may serve some biological

function, even if only to facilitate transcription of genes. RNA interference and gene silencing may also be induced by intergenic transcription, via siRNA mechanisms which induce heterochromatin formation. However again the more probable scenario for the majority of intergenic transcripts is that they are merely examples of background transcription with no biological significance. Further work would need to be done to determine where the transcripts are initiated as it may help explain why the retroelement is acting as a transcription stop site. While a large number of the recovered transcripts are examples of retroelement regulatory regions acting as termination signals, this suggests the retroelements are not methylated and therefore they are also likely to be capable of acting as transcription start sites.

It is commonly believed that retroelements are uniformly silent, except in the case of disease, when the mechanisms used to repress the retroelements are not in place (Yoder *et al.* 1997). Our analysis showed activity of retroelements to be as common in healthy tissue as in diseased tissue (see Figure 3-4). This finding suggests that there is a low level of transcription from retroelements which is occurring in normal healthy cells. *This again raises the question – does retroelement transcription have any biological significance or is it just background transcription?*

The analysis of chromosomal location of unique retroelement transcripts in the human genome highlights the fact that there is no real pattern to retroelement activation. The number of active retroelements showed a trend to decrease with chromosome size (with the exception of chromosome 19), but there was no obvious pattern which would predict which retroelement family could be expected to be activated. As discussed in Section 3.2.4, since chromosome 19 has a large number of active genes, the high number of active retroelement regulatory regions could be explained by chromatin changes induced by gene activation, leading to retroelement activation.

All chromosomes had retroelements which were active as either a transcription start or stop site, with the families of retroelements with the highest copy numbers having the most active LTRs/elements i.e. L1H elements are present in approximately 1000 copies, HERV-H are present in approximately 1306 copies and ERV-9 are present in approximately 418 copies. This suggests that a number of retroelements are activated due to local chromosomal changes. *In other words when a gene in the region of the retroelement is activated, the retroelement may become demethylated and active.* An example supporting this view is the IAP element located near the circadian gene m. nocturnin. Over time the IAP element becomes activated due to the repeated activation of the m.nocturnin gene.

If there was an obvious pattern of retroelement activation, such as with the HERV-F retroelement which is only expressed from one locus on chromosome 7, it might be speculated that the majority of retroelements are under some regulatory control and therefore may have biological significance. This does not appear to be the case, although small scale patterns may be lost in the overall picture when large numbers of retroelements are being examined. Our preliminary analysis however suggests that conclusions can not be drawn just by looking at positioning of active retroelements to predict which families will be activated on which chromosome. A more in depth analysis would be needed.

This study used strict guidelines for inclusion so transcripts were only included if they were obviously initiated/terminated within a regulatory region. It is highly likely that by doing so authentic transcripts were excluded, reducing the number of recovered transcripts by as much as 10 fold. Transcripts which had been truncated due to the RT enzyme falling off the template transcript before completing the run, and transcripts from families such as HERV-L which do not have a well defined

start site of transcription, are examples of transcripts which we excluded. Therefore this study should not be viewed as definitive, but rather as a guide as to the number of retroelements in the human and mouse genome which are active and therefore capable of transcription.

A problem with this type of study involves the EST sequences themselves. EST sequences and libraries do not always give a true indication of what is being transcribed in a cell or tissue. Different methods are used to produce the libraries – some being more comprehensive than others, and this creates problems when groups attempt to standardise their libraries. Normalisation, subtraction and size fractionation of libraries occur and occasionally retroelement sequences are removed as they are considered "junk". For these reasons many retroelement transcripts could have been missed. Additionally transcripts expressed at a low level are sometimes missed (and retroelements may fall into this category) due to the way a library is made. Perhaps one of the largest concerns though relates to contamination of libraries with DNA, pre-mRNA and other artefacts. Our study did not screen for this type of contamination as we did not have the ability to do so, but a recent paper by Sorek and Safer found that 53/6649 libraries they examined, were highly contaminated (Sorek and Safer 2003). Upon examining the libraries our ESTs were located in, there does not seem to be heavy weighting towards any of the listed libraries which are contaminated, but it certainly does not exclude the fact that some of our sequences may be artefacts which have resulted due to library contamination.

This study was conducted by examining <u>each individual element</u> returned from the BLAST searches "by eye" to ensure authentic transcripts were only included in the study. This meant the process was extremely time consuming, but it also means it was perhaps more precise than if a program was used to sort through the results.

In this chapter we have shown that retroelement transcription is common in both mice and humans, with a large number of retroelement families having intact regulatory regions capable of acting as transcription start/stop sites. While a large number of ESTs were recovered in both the mouse and human analysis, we realised that the transcripts needed to be authenticated to show they are real examples of transcription as there are many artefacts in the dbEST. This is when we became interested in running transposon display reactions, which could be used to obtain examples of legitimate retroelement transcription.

CHAPTER 4 TRANSPOSON DISPLAY

4.1 Introduction

The *in silico* analysis unearthed a number of transcripts which appeared to be initiated in retroelement promoters and were joined to cellular gene sequences. These chimeric transcripts did not appear to be present in large numbers in the EST libraries, but they were common enough to suggest that transcriptional interference via readthrough transcription by retroelements may not be limited to the few examples previously described.

Since we felt it was likely that many of the chimeras expressed in the EST libraries were peculiar to the individual mice or humans from which they were obtained, we decided to run transposon display experiments which would allow us to obtain chimeric candidates from individual mice that could be compared with other mice. Also, we could obtain a large amount of tissue from each mouse, so further experiments could be done to authenticate the transcripts.

The transposon display method was first described by Van den Broeck *et al.* (1998) where it was used to investigate the *dTph1* transposable element family in *Petunia hybrida*. Transposons are often responsible for variegation and mutation in plants; they appear to be less tightly controlled in plants than in mammals. Van den Broeck *et al.* (1998) devised the method in order to identify which particular transposon was causing a mutant phenotype in the W138 *Petunia hybrida* line (Van den Broeck *et al.* 1998).

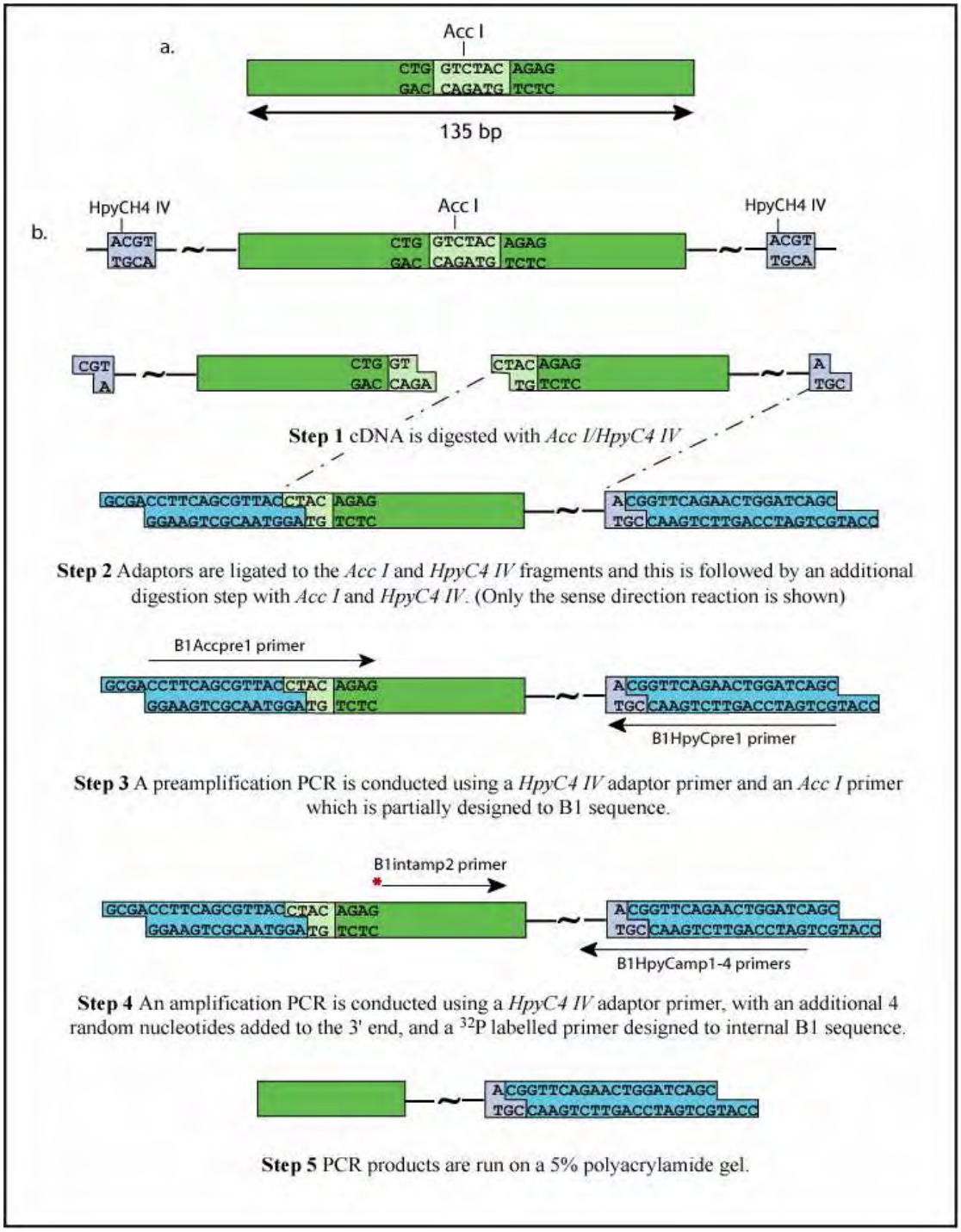
a.
Acc I
CTG GTCTAC AGAG
GAC CAGATG TCTC
135 bp
b.
HpyCH4 IV
ACGT
TGCA
Acc I
CTG GTCTAC AGAG
GAC CAGATG TCTC
HpyCH4 IV
ACGT
TGCA
CGT
A
CTG GT
GAC CAGA
CTAC AGAG
TG TCTC
A
TGC
Step 1 cDNA is digested with Acc I/HpyC4 IV
GCGACCTTCAGCGTTACCTAC AGAG
GGAAGTCGCAATGGATG TCTC
ACGGTTCAGAACTGGATCAGC
TGCCAAGTCTTGACCTAGTCGTACC
Step 2 Adaptors are ligated to the Acc I and HpyC4 IV fragments and this is followed by an additional digestion step with Acc I and HpyC4 IV. (Only the sense direction reaction is shown)
B1Accpre1 primer
GCGACCTTCAGCGTTACCTAC AGAG
GGAAGTCGCAATGGATG TCTC
ACGGTTCAGAACTGGATCAGC
TGCCAAGTCTTGACCTAGTCGTACC
B1HpyCpre1 primer
Step 3 A preamplification PCR is conducted using a HpyC4 IV adaptor primer and an Acc I primer which is partially designed to B1 sequence.
B1intamp2 primer
GCGACCTTCAGCGTTACCTAC AGAG
GGAAGTCGCAATGGATG TCTC
ACGGTTCAGAACTGGATCAGC
TGCCAAGTCTTGACCTAGTCGTACC
B1HpyCamp1-4 primers
Step 4 An amplification PCR is conducted using a HpyC4 IV adaptor primer, with an additional 4 random nucleotides added to the 3' end, and a 32P labelled primer designed to internal B1 sequence.
ACGGTTCAGAACTGGATCAGC
TGCCAAGTCTTGACCTAGTCGTACC
Step 5 PCR products are run on a 5% polyacrylamide gel.

Figure 4-1 Transposon Display method – adapted from Van den Broeck *et al.* 1998. a. Schematic representation of a B1 element from Mus musculus showing the restriction site for *Acc I* which is located approximately half way along the element. **b.** Schematic representation of the Transposon Display method.

Step 1. Depending on the enzymes used for the assay, either a double digestion or two separate digestion steps are carried out on the prepared cDNA. A tetracutter and a cutter which has a recognition site of 5 bp or more are used. In the case shown, *Acc I* (an enzyme with a 6 bp recognition site) cuts once within the B1 element and *HpyCH4 IV* does not cut within the B1 element so two fragments containing retroelement and flanking sequence are formed.

Step 2. Adaptor sequence is ligated to both *Acc I* and *HpyCH4 IV* fragments and ligation artefacts are then removed by carrying out an extra digestion step as described in Step1.

Step 3. A preamplification round of PCR is conducted using an adaptor primer to the *HpyCH4 IV* end and a primer designed to adaptor and retroelement at the *Acc I* end.

Step 4. A nested hot PCR is conducted with a primer which is designed to internal retroelement sequence, and which is labelled with ^{32}P γ-ATP, and an adaptor primer to the *HpyCH4 IV* end which has an additional 4 random nucleotides added to the 3' end.

Step 5. Samples are run on a 5% polyacrylamide gel so banding patterns may be compared. Bands of interest are excised, reamplified and sequenced.

Transposon Display is based on Amplified Fragment Length Polymorphism (AFLP) and reveals detailed information about copy number and insertion events of transposons. Since it uses DNA as the starting material, it does not provide information on transcriptional activity of repetitive elements such as retroelements; instead it permits assessment of the genetics of transposon variation.

In 2002 Kashkush *et al.* modified the Transposon Display to use cDNA as the starting material (refer to Figure 4-1 for a detailed outline of the method); and identified transcriptionally active Wis 2-1A retrotransposons in wheat. While other papers had used similar differential displays to look at both transposon and retrotransposon activity (Casa *et al.* 2000; Vinogradova *et al.* 2002), none had examined the possibility that active retroelements could be interfering with expression of surrounding genes. Kashkush *et al.* (2002) specifically addressed this question by examining chimeric Wis 2-1A elements and showing that certain elements were exerting transcriptional interference on nearby genes. This study was of particular interest to us, because we wanted to ask the same sorts of questions about the mouse.

Our dbEST search showed that there are a large number of mouse retroelements that have intact promoters and are capable of transcription. In addition, the presence of chimeric transcripts in the data analysed suggested that a large number of retroelements may be involved in transcriptional interference. However it was unclear how many chimeric transcripts were authentic examples of readthrough transcriptional interference.

The aim of this chapter is to,

- examine the activity of the IAP, MMTV, MuLV, VL30, B1 and L1Tf families of mouse retroelements using the transposon display method described by Kashkush *et al.* (2002).
- to determine what proportion of retroelement transcripts detected by the transposon display are chimeric and therefore may be causing transcriptional interference.

4.2 Results

In order to obtain a broad view of retroelement activity, six different retroelement families were examined using the transposon display; the ERV families IAP, MMTV, MuLV and VL30, the SINE family B1 and the LINE family L1Tf. These families were chosen because they are known to be active (refer to Chapter 3, Section 2.5) and so would be likely to produce bands in a transposon display.

The RNA for the experiments was taken from the liver or kidneys of either an a/a C57Bl/6 mouse, an A^{vy}/a C57Bl/6 mottled mouse or an A^{vy}/a C57Bl/6 yellow mouse. These mice were chosen due to the fact the display methods could be optimised by looking at expression of the IAP antisense promoter (see Section 4.2.1). Due to the fact the primary interest of the study was to identify chimeric retroelements and not categorise the elements by expression and sex/age parameters of the mice, male and female mice were used and the ages of the mice ranged from 3 weeks to 9 months. Tissue choice was random and was driven by the fact that the liver and kidneys could provide a large amount of RNA for the experiments. Each display was run on tissue taken from a single mouse - 1 a/a C57Bl/6 mouse, 1 A^{vy}/a C57Bl/6 mottled mouse and 1 A^{vy}/a C57Bl/6 yellow mouse.

The transposon displays were run as described by Kashkush *et al.* (2002) with some modifications. After the ligation, in the restriction/ligation step, an additional restriction round is included to remove any artefacts caused by religation of cut cDNA. In the preselective amplification step 6 µl of the restriction/ligation mix was added instead of 4 µl, to increase the amount of product being amplified. Finally, excised bands were cleaned up using a QIAex II kit. This was done to remove excess urea which was found to be interfering in the reamplification step.

Each transposon display uses a different set of primers and adaptors, due to sequence differences between families, and this meant each display needed to be optimised. Therefore it was at times necessary to run a display numerous times with different conditions and different components before a result could be obtained. This would not have been the case if we had analysed a smaller number of retroelements, but we conducted a large study in the hope of better understanding how often retroelements form chimeric transcripts.

The transposon displays produce a large amount of information; not only were a large number of bands amplified, but each band may also contain more than one type of chimeric sequence. To try and avoid missing candidates, 3 clones of each reamplified band were sequenced. Often the sequence for all 3 clones would be the same but on occasion, and particularly with the SINE elements, the sequences differed.

Not all transposon displays had every band excised, reamplified and sequenced, due to the amount of time it would take to carry this out for each of several hundred bands. Only the IAP antisense and B1 sense and antisense displays had all the bands examined. We did this so that we could identify how many of the bands in a typical display were likely to be artefacts versus unique transcripts. Artefacts were sequences which were amplified by the transposon display but which did not contain retroelement sequence, or they were sequences which had ligated to adapters incorrectly resulting in an incorrect sequence being amplified. Thus from the outset we had to restrict the analysis, and it should be clear that **the candidates we have recovered are only a sample (a random sample) of all the possible candidates**.

4.2.1 IAP Transposon Display

The transposon display method was optimized for mouse retroelements by examining the IΔ1 IAP family antisense promoter. RNAs used in the displays were

taken from the kidney of an isogenic strain that is mostly C57BL/6J in genetic background. In this strain the agouti locus is either "a" (nonagouti) or "A^{vy}" (agouti viable yellow), i.e. the strain is congenic at agouti. We analysed a/a mice, A^{vy}/a C57BL/6J mottled mice and A^{vy}/a C57BL/6J yellow mice: thus the IΔ1 IAP antisense display had an internal control – expression of the A^{vy} transcript, which originates on the IAP that is inserted into the agouti locus in the A^{vy} allele (Duhl *et al.* 1994). We would expect to find a band derived from the A^{vy} RNA transcript in the A^{vy}/a mottled and yellow mice, but not in the a/a mice.

There are a number of IAP subclasses, which differ from each other in the LTR region (Kuff and Lueders 1988) - this means a single display is only likely to identify a proportion of active IAP elements. This display was only designed to examine the IΔ1 IAP, which is known to be very active, and therefore results are not likely to be indicative of IAP elements in general. Refer to Figure 4-2 to observe differences in band patterns for the IAP antisense display. It should be noted that Figure 4-2 is not an accurate depiction of display banding pattern differences due to the low resolution of the picture. The film did not scan sufficiently well enough to show all bands in the low intensity range which gives the display an appearance of multiple expression differences between strains. Additionally the low intensity bands are likely to represent elements which are expressed at a low level which may account for the differences in band intensity between strains.

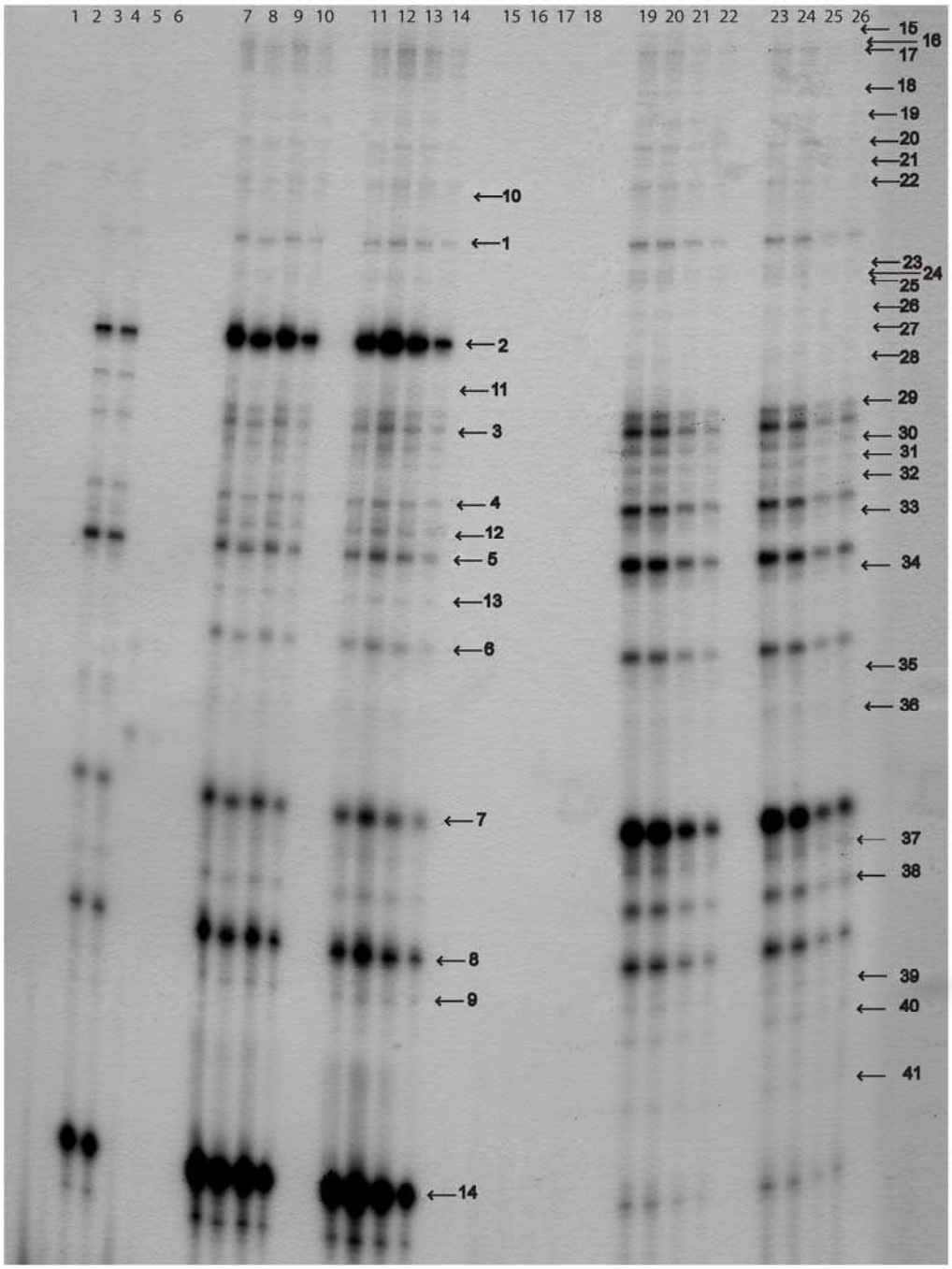
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26
15
16
17
18
19
20
21
22
10
1
23
24
25
26
27
2
28
11
29
3
30
31
32
4
33
12
5
34
13
6
35
36
7
37
38
8
39
9
40
41
14

Figure 4-2 IAP antisense cDNA transposon display. Lanes 1 and 2 A^{vy}/a C57BL/6J yellow –RT, 3 and 4 A^{vy}/a C57BL/6J yellow cDNA, 5 and 6 A^{vy}/a C57BL/6J mottled –RT, 7-14 A^{vy}/a C57BL/6J mottled cDNA, 15 and 16 water control, 17 and 18 a/a C57BL/6J –RT, 19-26 a/a C57BL/6J cDNA. Bands indicate possible chimeric retroelement candidates. The bold arrow indicates the A^{vy} IAP chimeric band which is present in the A^{vy}/a C57BL/6J mottled and yellow cDNA but absent in the a/a C57BL/6J cDNA. All other bands were excised, reamplified and sequenced to identify chimeric retroelement candidates. The display was carried out on cDNA made from the kidney.

The antisense IAP display produced multiple bands, although interestingly there were not many differences in the pattern of bands between the three cDNA samples. There was however one obvious band which differed between the samples; a dark band prominent in the A^{vy}/a cDNA lanes (both mottled and yellow) but absent in the a/a cDNA lanes. When the band was sequenced it proved to be the chimeric IAP/agouti transcript, indicating that the display had worked as it should.

In order to determine how many of the bands were chimeric retroelement candidates, all bands were excised, reamplified and sequenced. Of the chimeric candidates sequenced, one was found to be the CDK5 activator binding protein IAP chimera (Druker *et al.* 2004), three were chimeric with characterized genes (Copine 8, Non imprinted in Prader-Willi/Angelman syndrome 2 and Exocyst complex component 4), four candidates were found to be chimeric with hypothetical genes, eleven candidates were found to be located between genes, and one could not be located within the mouse genome (due to sequence similarity between retroelements). For diagrammatical representations of chimeric IAP examples refer to Figure 4-3. Refer to Table 4-1 for tabulated results.

We also found that some bands which ran at different sizes on the gel were actually the same chimeric retroelement. This was a common problem with all

subsequent displays and is likely to be due to DNA compression (Cherry 1992), where the secondary structure differs between two populations of molecules within the one PCR sample, making them run at different rates. The urea gel is intended to prevent this from happening, but clearly it is not absolutely effective.

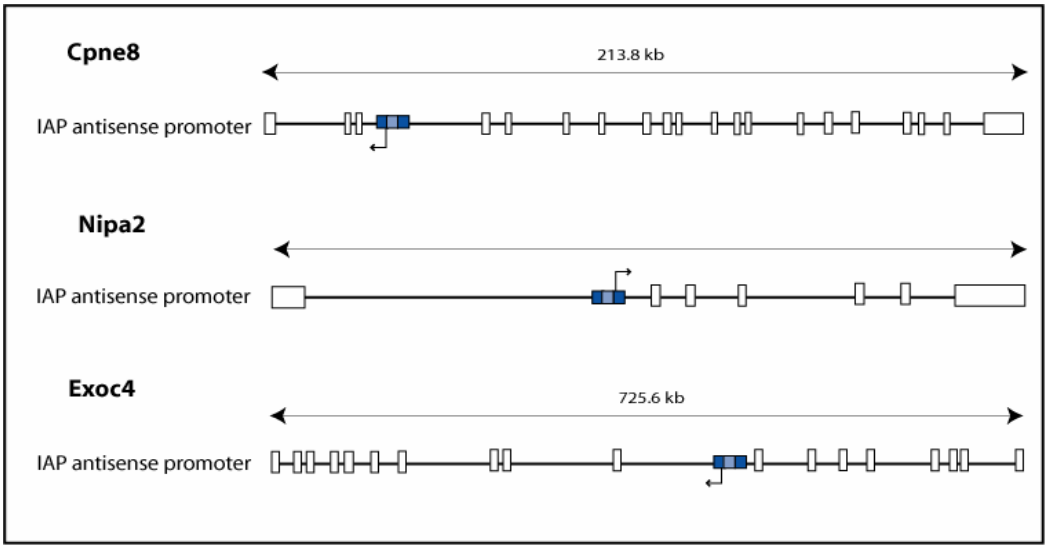

Figure 4-3 Locus diagrams for the Cpne8, Nipa2 and Exoc4 genes. The diagrams show the placement of the IAP element (blue box) in relation to the gene exons (white). The IAP antisense promoter direction is indicated.

Following the success of the antisense IAP display, we ran the transposon display for the IAP sense primers. In this case however there was a problem with contaminating bands appearing in the –RT (no reverse transcriptase) lanes. While the experiment was repeated numerous times with different conditions and new components, the bands could not be eradicated so the display could not be analysed.

4.2.2 SINE Transposon Display

The next retroelement family to be examined was the B1 SINE family of retroelements. B1 retroelements make up 2.66% of the mouse genome and are present

in 564 000 copies (Waterston *et al.* 2002), therefore it was expected that the display might produce a large number of bands.

While components such as the adaptors, enzymes and primers differed in the B1 display compared to the IAP display, both the B1 sense and antisense displays still produced a large number of bands. All the bands were excised, reamplified and sequenced in order to establish how many were chimeric candidates. The pattern of bands did not differ between the samples in the B1 sense reaction, nor did they differ in the B1 antisense reaction, *suggesting that all the mice expressed the same chimeric transcripts*.

In the B1 sense display twelve retroelement candidates were chimeric with introns of characterized genes: Betaine-homocysteine methyltransferase, Chemokine-like factor super family 8, COP9 homolog subunit 7b, Dual specificity phosphatase 22, Esterase 1, FK506 binding protein 12-rapamycin associated protein 1, Kinesin family member 5A, Hepsin, Nuclear receptor subfamily 6 group A member1, Platelet-activating factor acetylhydrolase isoform1b beta 1 subunit, PR domain containing 2 with ZNF domain and RAB10. For diagrammatic examples refer to Figure 4-4. An additional seven candidates were chimeric with exons of genes – all of these exonic candidates were located within the 3' or 5' untranslated regions of the gene sequence. Six chimeric retroelement candidates were chimeric with hypothetical genes and eight chimeric candidates were found to be in intergenic regions.

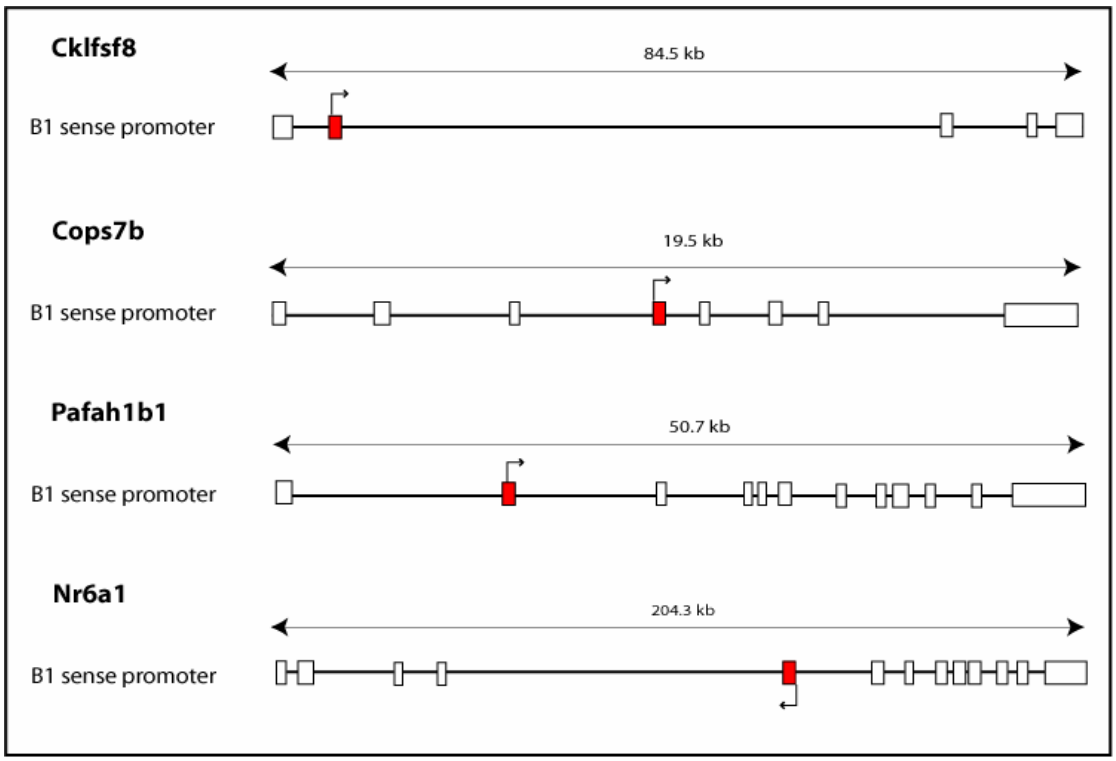

Figure 4-4 Locus diagrams for the Cklfsf8, Cops7b, Pafah1b1 and Nr6a1 genes. The diagrams show the placement of the B1 element (red box) in relation to the gene exons (white). The B1 sense promoter direction is indicated.

The B1 antisense display likewise produced a large number of chimeric retroelement candidates which were chimeric with intronic sequence of characterized genes: Ariadne homolog 2, ATPase family AAA domain containing 3A, Autocrine motility factor receptor, Cyclin-dependent kinase 7, Esterase 1, Glucosamine (N-acetyl)-6-sulfatase, Nuclear receptor coactivator 1, Phosphoinositide-3-kinase adaptor protein 1, Protein-L-aspartate O-methyltransferase 1, Rab31 like, Rrn3, ST3 beta-galactoside alpha-2 3-sialyltransferase 4, Transforming acidic coiled-coil containing protein 2 and 1-acylglycerol-3-phosphatse O-acetyltransferase 2. For schematic examples of gene loci refer to Figure 4-5. Five candidates were found to be chimeric

with gene exons, three candidates were chimeric with hypothetical genes, one candidate was located between genes, and two could not be located within the mouse genome. Refer to Table 4-1 for tabulated results.

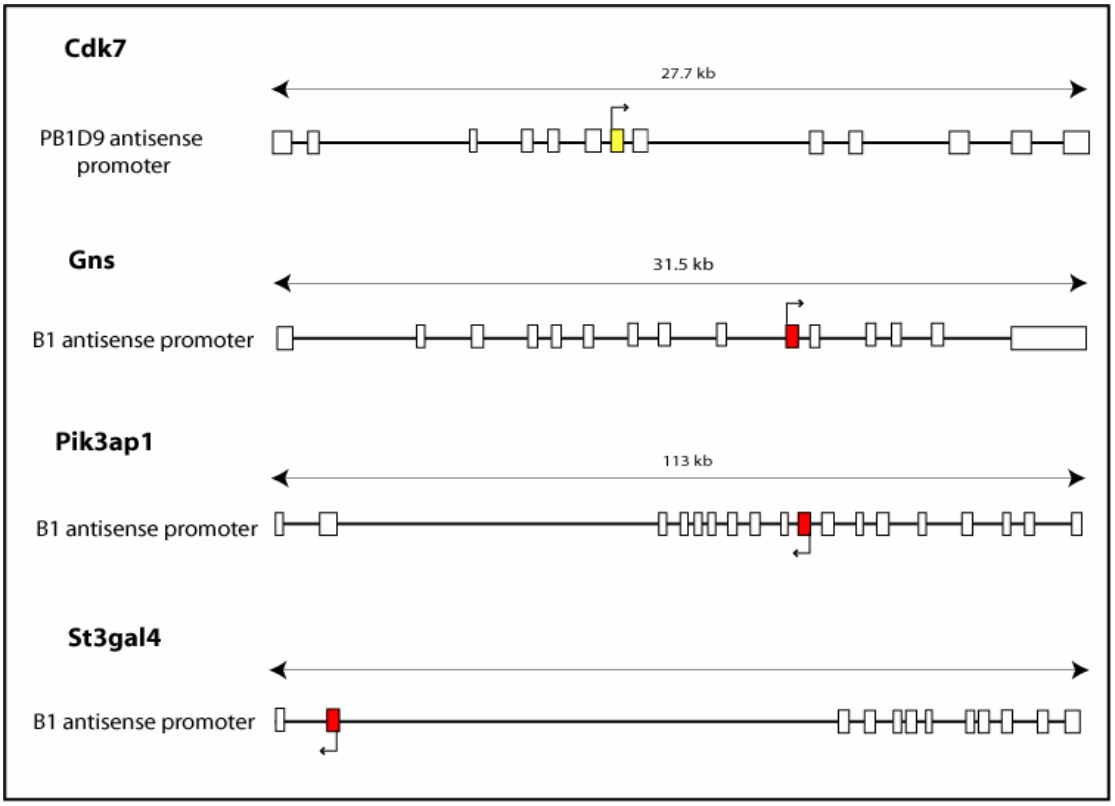

Figure 4-5 Locus diagrams for the Cdk7, Gns, Pik3ap1 and St3gal4 genes. The diagrams show the placement of the B1 element (red box) and the SINE element PB1D9 (yellow box) in relation to the gene exons (white). The B1 or PB1D9 antisense promoter direction is indicated.

Results from both the sense and antisense B1 display supported the dbEST analysis, suggesting that a large number of SINE promoters are intact and capable of transcription.

4.2.3 MMTV Transposon Display

The MMTV retroelement family is present in approximately 10 copies in the mouse genome (Boeke and Stoye 1997). This is a very small number compared with the other ERV families examined in this study, which have hundreds of copies per mouse genome. Since there are so few copies of MMTV, it was expected that the number of chimeric transcripts would be correspondingly small.

The display for the MMTV sense primers produced fewer bands than the B1 and IAP displays, but more than would be expected for a retroelement family with only 10 copies per genome. Sequencing demonstrated that nearly all the bands were artefacts, sequences which can be amplified by the primers but are not within a retroelement. Artefacts had been detected in the IAP and B1 displays, but in low frequencies compared to the MMTV display.

The MMTV antisense display initially did not produce any bands, and several attempts were made to recover candidates by changing the primers. Eventually with one set of primers a number of bands were amplified (refer to Figure 4-6). All were artefacts similar to those recovered from the sense display.

Only twelve bands were sequenced from each MMTV display and only two MMTV sense candidates were obtained. Unfortunately the location of the retroelements within the genome could not be determined so they were not further analysed. Refer to Table 4-1 for results.

It is possible more examples of chimeric retroelement candidates may be found if primers were redesigned, but this was not attempted because it would require a large amount of time and the outcome was uncertain.

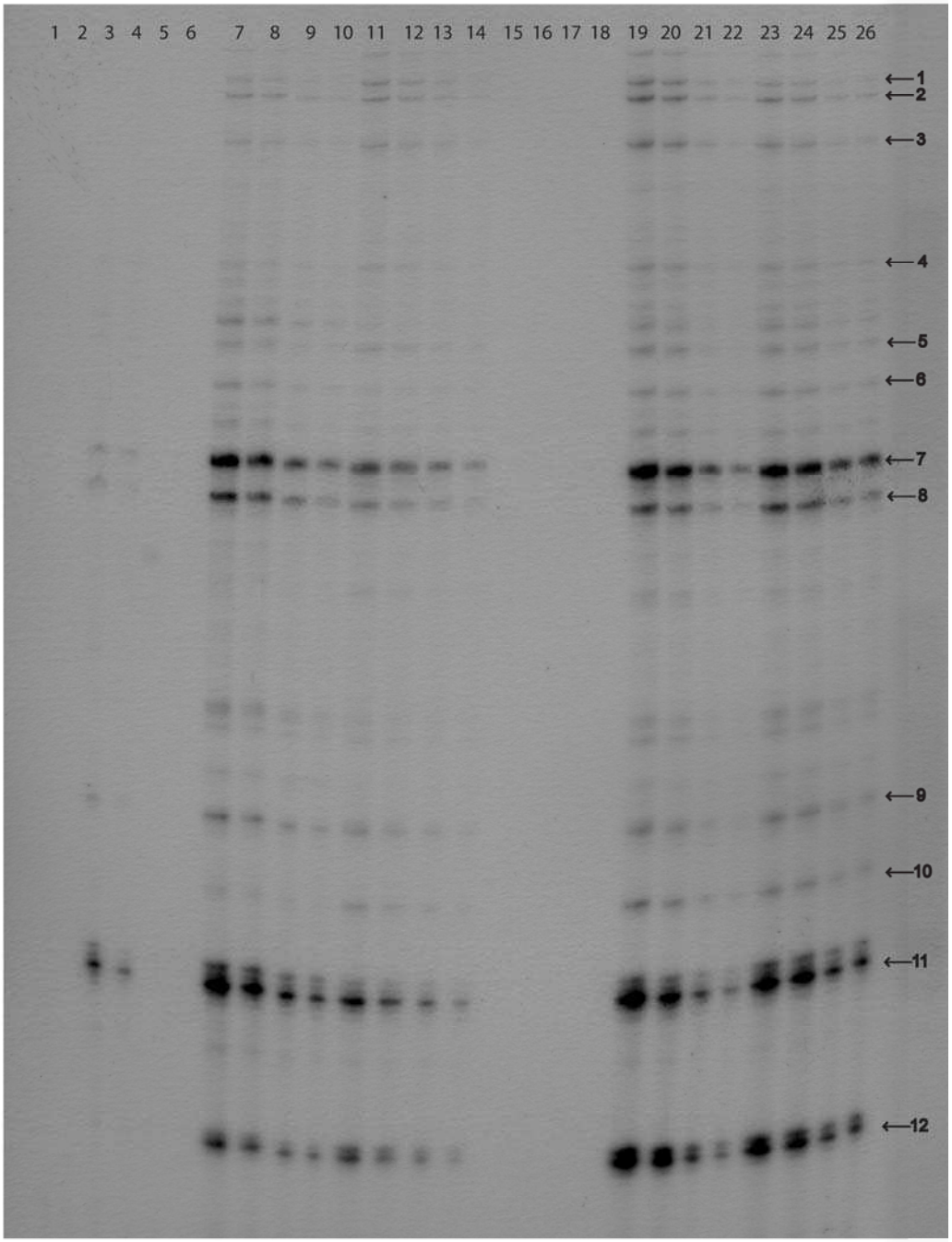
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26
1
2
3
4
5
6
7
8
9
10
11
12

Figure 4-6 MMTV antisense cDNA transposon display. Lanes 1 and 2 A^{vy}/a C57BL/6J yellow –RT, 3 and 4 A^{vy}/a C57BL/6J yellow cDNA, 5 and 6 A^{vy}/a C57BL/6J mottled –RT, 7-14 A^{vy}/a C57BL/6J mottled cDNA, 15 and 16 water control, 17 and 18 a/a C57BL/6J –RT, 19-26 a/a C57BL/6J cDNA. The display was carried out on cDNA made from the liver. Bands indicate possible chimeric retroelement candidates. A number of faint bands were amplified suggesting artefacts had been amplified since there are only 10 copies of MMTV per genome. 12 bands were excised, reamplified and sequenced to identify chimeric retroelement candidates.

4.2.4 MuLV Transposon Display

Initially when the sense and antisense MuLV displays were run, no bands were amplified. This seemed strange because there are approximately 25-100 copies of MuLV in the mouse genome (Boeke and Stoye 1997). Using new primers and adaptors, a number of bands were amplified.

The MuLV family has a well characterised LTR region, and primers were designed to a region conserved between the various subgroups of MuLV. Both the sense and antisense displays produced a number of bands, but the unique sequence portions of many of these candidates were very short (less than 50 bp) or the unique portion could not be located in the genome, which meant nearly all the elements sequenced could not be located in the genome. The MuLV family is highly conserved, and so the retroelement portion of the chimeras could not be used to locate the transcript in the genome. Together these problems prevented us from further analysing the chimeras.

For the MuLV sense display one chimeric candidate was sequenced which contained MuLV sequence spliced to exon 14 of the gene Annexin 6 (refer to Figure 4-7). Since this is a coding exon, this transcript provides direct evidence of transcriptional interference by a retroelement. Additional chimeric candidates were also

sequenced – one was chimeric with a hypothetical gene, while the remainder could not be located in the mouse genome.

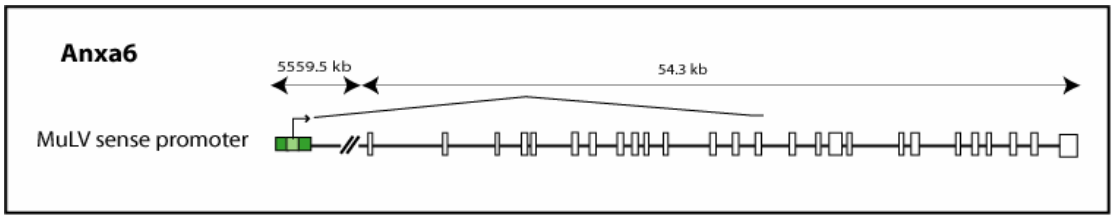

Figure 4-7 Locus diagram for the Anxa6 gene. The diagram shows the placement of the MuLV element (green box) in relation to the gene exons (white). The MuLV sense promoter direction is indicated.

The MuLV antisense display produced one candidate which was located within intronic sequence from the Calcium/calmodulin-dependent protein kinase II gene. The candidate contained only retroelement sequence, which was however sufficiently different from other MuLVs to locate it in the genome. Two chimeric candidates were found to be between genes (these may be considered "intergenic transcripts") while most of the candidates could not be located in the genome. Refer to Table 4-1 for tabulated results.

As was observed in the MMTV displays, new primers could have been designed in the attempt to locate where the active elements are within the mouse genome. However since the experiment had already been run with two different sets of primers, it seemed likely again that for the time invested, the results returned would not have been rewarding.

4.2.5 VL30 Transposon Display

There are 100-200 copies of VL30 in the mouse genome (Boeke and Stoye 1997) and results from dbEST analysis suggested that a number of these have intact promoters which are capable of transcription.

While a large number of bands were amplified in the VL30 sense display, many of the recovered sequences could not be located in the mouse genome - due to the small amount of unique sequence present. Of the bands sequenced, one candidate was chimeric with intronic sequence from the Rho guanine nucleotide exchange factor 12, one was chimeric with a hypothetical gene, and eight candidates reside in intergenic regions. Refer to Figure 4-8 for a diagram of the Arhgef12 locus.

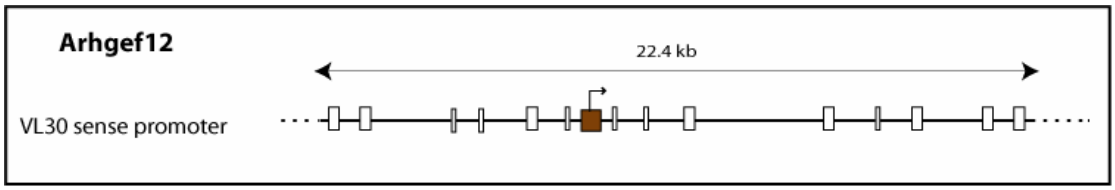

Figure 4-8 Locus diagram for the Arhgef12 gene. The diagram shows the placement of the VL30 element (brown box) in relation to the gene exons (white). The VL30 sense promoter direction is indicated. Only some of the Arhgef12 exons are depicted due to the large number of exons in the gene; the dashed lines indicate where the extra exons are placed.

The VL30 antisense display initially produced a large number of artefacts, so it was rerun twice with new primers and adaptors. The third run still produced some artefacts, but the numbers were reduced. Unfortunately the recovered sequences produced only one candidate and it was from an intergenic region. The rest could not be located, due to the high sequence similarity between family members and the small amount of unique sequence amplified. Refer to Table 4-1 for results.

Sequences of VL30 retroelement LTRs are quite variable (computer analysis data not shown) which indicates that the display is likely to have examined only a small subset of active VL30 elements. Further transposon displays with different primers would be likely to produce more chimeric candidates.

4.2.6 L1Tf Transposon Display

The L1Tf family of LINEs contains about 3000 full-length members per diploid mouse genome (DeBerardinis *et al.* 1998; Naas *et al.* 1998) with members having 99% sequence similarity. As with other murine LINE elements, the 5' end of the L1Tf sequence contains a variable number of tandemly repeated units of 205-210 bp called monomers. This was a problem when designing primers for the antisense display, as it meant the same active element could be amplified multiple times depending upon which monomer the primer bound to. This problem could not be overcome, so the display was run with a primer designed to hybridise to sequence located towards the start of the monomer sequence.

The initial runs of the L1Tf displays produced no bands. When the method was modified so that the preamplification primers lacking the 4 random nucleotides at the 3' end were used in the amplification step (see Figure 4-1), numerous bands were recovered. This may have occurred because the degeneracy at the 3' end of the primer restricted the number of monomers that could hybridize with the degenerate pool, or because the degenerate primers were being depleted during the PCR run by binding to many other L1Tf elements (there are approximately 3000).

A number of bands were amplified and sequenced in the L1Tf antisense display, but only three proved to be chimeric transcripts – L1Tf sequence joined to Unc-13 homolog B and to CUG triplet repeat RNA binding protein 2, and an L1Tf sequence spliced to exon 5 of Hydroxyacid oxidase 1. The latter is likely to be a good candidate

for transcriptional interference, because splicing to an exonic region has occurred. Refer to Figure 4-9 for locus diagrams.

Of the other bands sequenced, two candidates were chimeric with hypothetical genes, one was located in an intergenic region, and two could not be located in the mouse genome because insufficient unique sequence was amplified.

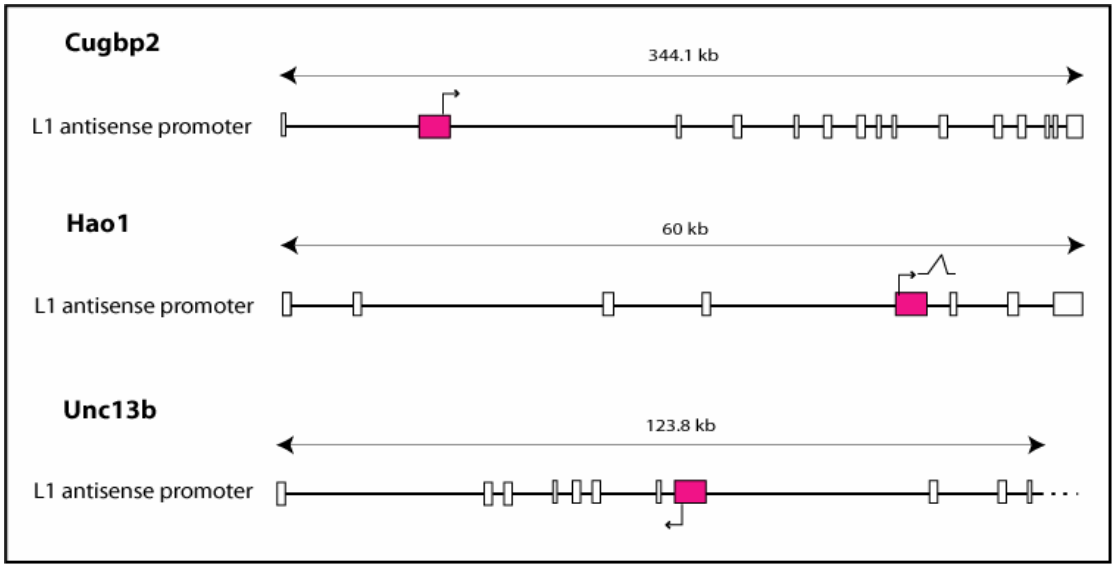

Figure 4-9 Locus diagram for the Cugbp2, Hao1 and Unc13b genes. The diagram shows the placement of the L1 element (pink box) in relation to the gene exons (white). The L1 antisense promoter direction is indicated. Only some of the Unc13b exons are depicted due to the large number of exons in the gene; the dashed lines indicate where the extra exons are placed.

The L1Tf sense display produced a large number of bands, but most were artefacts. One chimeric candidate was spliced to exon 4 of the Sorbitol dehydrogenase 1 gene making it a good candidate for further examination; one candidate was joined to intronic sequence of Plexin domain containing gene and one candidate was chimeric with a hypothetical gene. Refer Figure 4-10 for locus diagrams and to Table 4-1 for collated results.

Only twelve bands were sequenced for both L1Tf displays, so it is very likely that other candidates would be discovered by further sequencing.

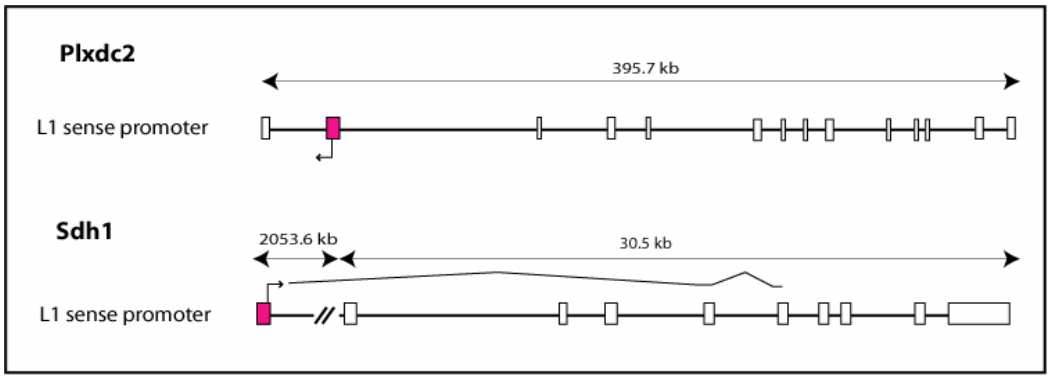

Figure 4-10 Locus diagram for the Plxdc2 and Sdh1 genes. The diagram shows the placement of the L1 element (pink box) in relation to the gene exons (white). The L1 sense promoter direction is indicated. The L1 element chimeric with the Sdh1 gene may be incorrect as multiple L1 elements were located in the BLAST search of the mouse genome. The L1 included in the diagram was the closest element to the Sdh1 gene by a large distance.

4.2.7 Transposon Display Results

Chimeric candidates obtained from the transposon displays of the six retroelement families are collated in Table 4-1.

Promoter	Retroelement within a gene	Retroelement within hypothetical gene	Retroelement between genes	Unknown location	Total no. of excised bands	Chimeric Candidate Examples
B1 sense	19	6	8	-	25	Hepsin, Chemokine-like factor super family 8
B1 antisense	19	3	1	2	20	Cyclin dependent kinase 7, Interferon gamma receptor 2
MMTV sense	-	-	-	2	12	-
MMTV antisense	-	-	-	-	12	-
MuLV sense	1	1	-	~6	12	Annexin 6
MuLV antisense	1	-	2	many	28	Calcium/calmodulin-dependent protein kinase II
VL30 sense	1	1	8	~4	22	Rho guanine nucleotide exchange factor 12
VL30 antisense	-	-	1	4	12	-
L1Tf sense	2	1	-	-	12	Sorbitol dehydrogenase 1
L1Tf antisense	2	2	1	2	12	Hydroxyacid oxidase 1 (liver)
IAP sense	n/a	n/a	n/a	n/a	n/a	n/a
IAP antisense	5	5	11	1	39	AvyIAP, Non-imprinted in Prader-Willi Angelman syndrome 2

Table 4-1 Chimeric candidates for the B1, MMTV, MuLV, VL30, L1Tf and IAP families.

4.3 Discussion

The genome is densely populated with retroelements that retain intact promoters, and so have the potential to transcribe either themselves or other sequences flanking them. However little information is available about the frequency with which chimeric retroelement-gene transcripts are formed. A small number of examples have been described in the mouse (Duhl *et al.* 1994; Flood and Ruvinsky 2001; Druker *et al.* 2004), but in all of these examples the transcript originates in the same type of retroelement, an IAP; with the exception of Speek's L1 examples from dbEST there is almost no data to suggest that other retroelement families are capable of transcribing adjacent genes. The data obtained from our dbEST analysis suggests that most retroelement families are capable of producing chimeric transcripts, and that chimeric transcript formation may actually be fairly common. Rather than settle for analysis of the transcripts that happen to have been placed in dbEST, we used the transposon display method described by Kashkush *et al.* (2002) to examine chimeric transcript formation in six active retroelement families: B1, IAP, MMTV, MuLV, VL30 and L1Tf.

The transposon displays produced chimeric candidates from nearly all families examined; five of the six families produced a large number of candidates for transcriptional interference, and only MMTV (present in a low number of copies) produced none. **Both antisense and sense transposon displays produced chimeras** which supports what we observed in the *in silico* analysis. As mentioned in Chapter 3 antisense transcription has been found to be a common occurrence in the human genome, although its function is unknown. The total number of unique chimeric candidates (which were located within the genome and therefore shown to be unique) for all the families examined was 100. This is a large number considering the

exploratory nature of the experiment. Approximately 70% of the clones we sequenced were chimeric transcripts (data not shown). The high percentage indicates that the transposon display method worked well and amplified a larger number of retroelement chimeras than artefacts.

The IAP and B1 families produced the largest numbers of chimeric transcripts, while the MMTV family produced the smallest number (refer to Table 4-1). To some extent this result is due to a bias in the number of bands that were cloned and sequenced: more bands were analysed in the IAP experiment, and so more chimeric transcripts were identified. The total number of bands excised and analysed for the B1 sense and B1 antisense displays were similar to the VL30 sense and MuLV antisense displays, and so the large number of candidates recovered from the B1 family displays is not due solely to a larger number of bands picked for sequencing. One explanation for the large number of B1 transcripts recovered is that these elements are small (~250 bp) and may therefore be tolerated in active gene regions where larger elements would be detrimental. Perhaps a more likely explanation is that there are 564 000 B1 elements in the genome, by far the largest number of all the families we examined; if all retroelements are equally likely to achieve an active state, then one would expect to find more chimeric transcripts originating in B1s. Conversely we would expect the MMTV family, with only approximately 10 members per genome, to produce few transcripts, as was observed.

Chimeric candidates were grouped into four main types based on their location within the genome (refer to Table 4-1 for details). Obviously the candidates of most interest are those chimeric with a known gene; these are the candidates which will be analysed further. This does not mean that the other examples are not equally as likely to be candidates for transcriptional interference, but rather they will be harder to find

information on and therefore analyse. Candidates chimeric with hypothetical genes could be analysed at a later time when more information is gathered about the genes, while the A^{vy} IAP is an example of how important intergenic chimeras can be.

One problem, which reduced the number of chimeric candidates obtained, was the issue of retroelement sequence similarity amongst family members. This was most obvious in the MuLV and VL30 displays. Primers for the MuLV and VL30 displays were designed in a region that appeared to be quite conserved amongst family members. Unfortunately a population of both MuLV and VL30 elements had an additional restriction site which was not present on the master sequences used to design the assays. This resulted in some cases with just retroelement sequence being amplified and in other cases with only a very small amount of unique sequence being amplified. Since the retroelement sequence amplified was highly conserved, it resulted in a number of sequences which could not be located within the genome and therefore which could not be further analysed.

While we attempted to design primers in conserved sequence regions of the retroelement families studied, due to the dependence of this method on primers, adaptors and experimental conditions and also due to the fact the primers can not be specific for every retroelement in a particular family in the genome, **the transposon displays are highly likely to have missed many active retroelement members.** The point of running the transposon displays in the mouse was to ask how frequently chimeric transcription occurred and not to establish a definitive number of retroelements which are forming chimeric transcripts. **The results we obtained can therefore be only a guideline for the numbers of retroelements which are capable of readthrough transcriptional interference.** Judging by the numbers we observed it is

likely that there are large numbers of retroelements which may be capable of transcriptional interference.

Not all the candidates obtained from the displays would be authentic examples of readthrough transcription due to the fact that the transposon display method has the potential to amplify a high number of artefacts. The display is designed in such a way that retroelements which have been incorporated into a cellular gene as the promoter (Landry *et al.* 2001; Dunn *et al.* 2003; van de Lagemaat *et al.* 2003) or stop signal (Britten 1997; Mager *et al.* 1999), will still be amplified in the display. These types of retroelements are no longer considered to be autonomous retroelements and are instead considered to be part of the cellular gene. In addition splicing and polyadenylation of genes is unlikely to occur in a perfect stepwise manner and it is therefore possible that some transcripts may be polyadenylated before splicing has finished. **Therefore some chimeric examples may be retroelements within intronic regions of RNA that were in the process of being spliced out**. This would be a plausible explanation for chimeric candidates which were found in only one of the mouse cDNA samples. But since nearly all the chimeric candidates were present in all three samples examined, this is an unlikely explanation for the occurrence of a large number of the chimeric candidates examined.

There have also been reports of some functional human genes being entirely derived from mobile elements - including repetitive elements (Britten 2004). Upon a brief analysis of a subset of chimeras, only one example contained a fairly high percentage of retroelement sequence (30%) however it was only a hypothetical gene. None of the 33 chimeras which contained retroelement and intronic sequence from characterised genes contained notable amounts of retroelement sequence within the coding regions. Therefore while some of the chimeric candidates (i.e. hypothetical

genes which are not completely characterised) may be similar murine examples of functional genes containing large proportions of mobile elements, it seems unlikely that many of the characterised gene chimeric examples would fall into this category. Since the largest number of chimeric candidates were from characterised genes, it strengthens the idea that many of the transcripts are likely to be authentic.

As mentioned earlier some artefacts will purely be due to adaptor errors and priming errors but on the whole this category of artefact had the lowest numbers and therefore did not significantly contribute to the number of artefacts recovered.

All the families of retroelements studied produced chimeric transcripts, *suggesting that most if not all retroelement families are capable of forming chimeric transcripts* - and thus exerting transcriptional interference on surrounding genes. Therefore it is possible that there are a large number of genes which form chimeras with retroelements. This adds another layer of complexity when we consider the way a genome is controlled and expressed. The variable activity of these retroelements may result in variable effects on gene expression as is seen with the A^{vy} and $Axin^{Fu}$ mice. **Transcriptional interference thus provides another explanation for diseases which cannot be explained by genetics and environment alone**.

Genes that were found to be chimeric with retroelements varied in function and cellular localisation, so they could not be placed into groups based upon similarities. The only similarity between the chimeras was the expression of the retroelements. This suggests that the activation of the retroelements is a random process with elements from all over the genome being affected. This results in different types of genes forming chimeras with the retroelements, which means they cannot be grouped on relatedness.

Only a small subset of chimeras was identified in only two types of mouse tissue (liver and kidney), so there is likely to be a large number of chimeras which were

missed. This means that a pattern for genes capable of forming chimeras might emerge upon further examination; however we think it is likely that the random nature of the activation of retroelements means that this will not occur.

An interesting observation from the transposon displays was that chimeric transcription did not substantially differ between the three cDNA samples. In other words, when the amplified samples were run on a gel, the patterns of bands were basically the same between the samples. This finding was inconsistent with our idea that retroelements are expressed in a different variegated pattern in each individual, resulting in different patterns of transcriptional interference between individuals (Whitelaw and Martin 2001). Our finding does not refute this hypothesis, in that there may be a core number of retroelements which are active in all individuals and it is only a small number which differ between individuals. It may be this small number which produce the phenotypic variation as is seen with the A^{vy} IAP. It is significant however that in the large number of retroelement chimeras which were sequenced, few differed between mice. Perhaps further displays would produce candidates which were unique to individual mice. Alternatively the unique transcripts may be expressed at a low level, compared to the commonly expressed transcripts, and they were just missed during band excision and reamplification.

We know inheritance of retroelement expression patterns can occur, as it has been shown to occur with the A^{vy} and $Axin^{Fu}$ IAP elements. Therefore it is possible that expression patterns of other retroelement chimeras may be inherited. Because the chimeric transcripts we recovered were present in all mice in the same pattern, we could not examine inheritance of retroelement expression, as this would require a differentially expressed retroelement. Quantitative PCR could have been used to examine the candidates we recovered, as there may be some concealed variation in the

expression of the chimeras. But this would have required a large amount of time to design and optimise conditions just to examine a few candidates, and we decided it would be a better use of time to run experiments such as primer extension and RT-PCR, which could show that the chimeras were authentic.

The patterns of expression observed for the retroelement chimeras also raised a question: **why should chimeric retroelements be conserved**? Before this question could be examined we needed to show the chimeric candidates that had been recovered from the transposon displays were actually authentic transcripts and were not just artefacts.

This chapter has shown that retroelement chimeric transcription is widespread in the mouse genome as five of the six families studied were capable of forming chimeric transcripts. To enable us to draw stronger conclusions, we next turned our attention to showing that the candidates, which had been amplified from the transposon displays, were authentic chimeric transcripts which had been initiated in the promoter of the retroelement.

CHAPTER 5 TRANSCRIPT AUTHENTICATION

5.1 Introduction

The transposon display provided us with a large number of chimeric retroelement candidates, which were possibly exerting transcriptional interference on surrounding genes via the mechanism of readthrough transcription. While a large number of chimeras were recovered, we could not be sure they were authentic transcripts unless further experiments were conducted. This chapter aims to authenticate retroelement chimeras by conducting primer extension reactions to identify the site of transcription initiation. Since retroelement activity state can be inherited (as observed with the A^{vy} and $Axin^{Fu}$ mice), we were also interested in seeing how common it was for retroelement chimeras to be differentially expressed. This chapter therefore also aims at using RT-PCR to establish whether expression of a random subset of retroelements varies between mice of different strains, as well as mice of the same strain.

The large volume of numbers returned by the transposon display, meant we had to choose between examining either a handful of chimeras in detail, or a larger number in less detail. We decided the latter would be more beneficial for two main reasons. If we looked at only a handful of chimeras in detail, it could mean that a large amount of time would be spent looking at randomly chosen specimens, perhaps only to find they were not authentic. Also it would not give a true indication of whether the transposon display was a good method to use to identify potential retroelement chimeras. Therefore we looked at a larger number of candidates, with the aim of providing a number of candidates which can be examined in more detail by further work.

In order to demonstrate that transcriptional interference is occurring, it is important to show that the transcription from the chimeric candidates is being initiated in the promoter of the retroelement (Whitelaw and Martin 2001). There were two

obvious methods which could be used to gain this information – primer extension or 5' RACE. **Primer extension** is an old method which is very robust and which has been shown to be quite sensitive, i.e. it can determine a promoter region when only small amounts of transcript are available. An advantage of this method is that it does not need optimising, as standard conditions are suitable for different primers; thus it is relatively easy to run a number of experiments in a short time. The main disadvantage is that the precise location of the promoter cannot be established – just a small but general region. **5' RACE** is a newer method which is very sensitive, since it is PCR based, so it can detect promoter start sites from much less starting material than a primer extension. It also has the advantage that the product can be cloned and sequenced to establish the precise position of the promoter. Its major disadvantages are that it requires optimising for each experiment, due to differences in primers, and compared to the primer extension it takes longer to look at a large number of samples. Since this was only a preliminary study aimed at looking at a large number of samples, we felt primer extension would be the best experiment to run in the given time.

In primer extension reactions the start site of transcription is mapped by using Reverse Transcriptase (RT) to extend sequence from a specific primer, to the end of the transcript (Refer to Figure 5-1). A 40 bp primer is designed to sequence which is approximately 100bp from the estimated start site of transcription of the gene, or in our case, the retroelement of interest. The 100 bp distance is selected in order to ensure the RT enzyme does not disassociate from the transcript template due to processivity constraints. The end labelled primer is added to total RNA, and allowed to slowly anneal to the template transcript before the RT is added. The RT binds to the primer and extends the sequence until it reaches the 5' end of the transcript, where it falls off the template transcript. A labelled newly synthesized transcript of a certain size

remains. The samples are then run on a 5% polyacrylamide gel and bands are produced which correlate to the length from the primer to the start site of transcription. The bands range in intensity depending upon the amount of transcript which was present in the total RNA. The reaction is deemed successful if a band corresponding to the expected product size is visualised.

One disadvantage of the primer extension is that it is quite common to recover multiple bands of different intensities in a given reaction. This results from the reverse transcriptase pausing or terminating transcription, due to secondary structure or extensive GC rich stretches in the RNA template (Triezenberg 1992). This premature termination of the transcript results in "weak" stops which are bands of low intensity. A "strong" stop (a dark band) generally indicates a possible promoter site. Bands of unexpected sizes may also be due to alternative promoters, so visualising more than one band does not indicate that the method has not worked. Thus interpretation of primer extension reactions is not always straightforward: if a single strong stop is present, then it is likely to represent the true promoter, but when multiple bands are present it can be difficult to establish which of them (if any) is the true start site. It helps to have a predicted start site, which is available for most retroelements.

While the transposon displays produced a large number of chimeric candidates, they were examined in the congenic A^{vy} mice (a/a, A^{vy}/a mottled and A^{vy}/a yellow). Since mouse strains are genotypically quite different to each other (Beck *et al.* 2000), we were interested in observing if other mouse strains express the same chimeric candidates. Results from this type of analysis would indicate whether the chimeras are strain specific, or whether they have been inherited from a common ancestor, an issue which has not been examined to date by anyone.

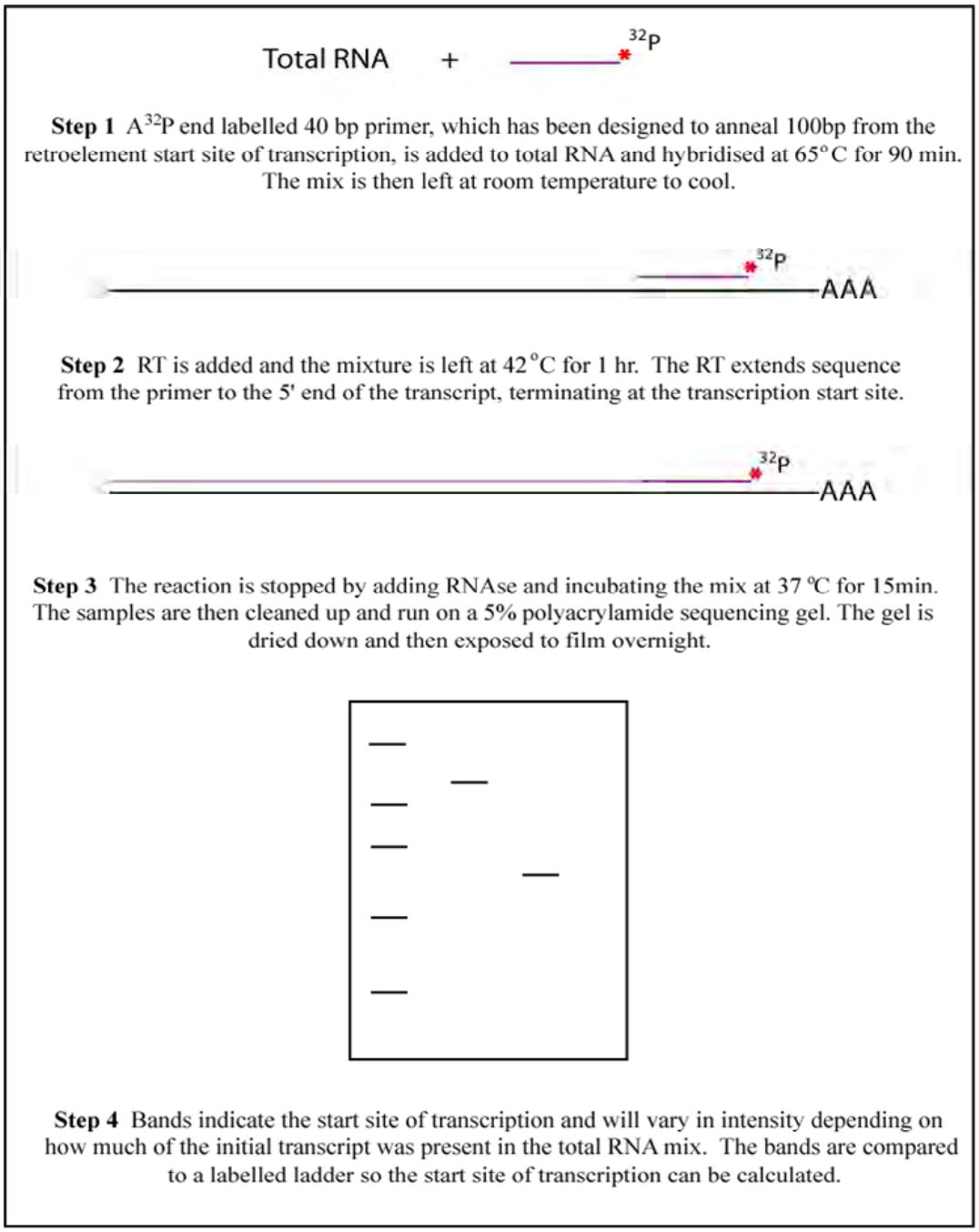

Figure 5-1 Primer Extension method.

A method which is widely used to show transcripts are present in an RNA pool is RT-PCR. It is a fast and easy method to use so we decided to examine chimeric candidate expression in different mouse strains via RT-PCR.

This chapter aims to,

- demonstrate that chimeric candidates are authentic transcripts by using primer extension reactions to show the chimeric transcripts are being initiated in the retroelement promoters.
- examine expression of the chimeric candidates in other mouse strains to observe whether expression of the chimeras is confined to one mouse strain or whether it is a widespread phenomenon.

5.2 Results

A large number of chimeric candidates were recovered from the transposon display experiments, and due to the limited time available only a selection of candidates could be examined further to test whether they were authentic transcripts. The candidates which were chosen for further examination were taken from three groups: transcripts in which a retroelement was spliced to an exon of a cellular gene (Anxa6, Hao1 and Sdh1), transcripts with chimeric retroelement-cellular gene intronic sequences (Agpat2, Amfr, Arhgef12, Arih, Atad3, Cdk7, Cklfsf8, Cops7b, Cpne8, Cugbp2, Es1, Frap1, Gns, Kif5a, Ncoa1, Nipa2, Nr6a1, Pafah1b1, Pcmt1, Pik3ap1, Plxdc2, Rrn3, Exoc4, St3gal4 and Unc13a), and transcripts which were chimeric with retroelement and exonic cellular gene sequence (Abcd4, Car5a, Ifngr2 and Smp1). The first two groups of candidates seemed to be the most likely to yield interesting results; they can easily be examined by primer extension and RT-PCR and are of more interest than a retroelement chimeric with a hypothetical gene, or a retroelement which is transcribing within an intergenic region. Examining the exonic chimeras may be harder, as observing a difference between the gene-driven transcription and transcription coming from the retroelement promoter may be difficult. Again it should be stressed that all the candidates recovered from the display are likely to be candidates for transcriptional interference, and with unlimited time all could be studied in greater detail.

As was mentioned previously, the computer analysis also uncovered a number of chimeric retroelement transcripts (transcripts which have been initiated in the promoter of a retroelement and which are joined to cellular gene sequence). We wanted to know if expression of a selection of these chimeras was widespread amongst different mouse strains. This resulted in three candidates being randomly chosen for further analysis via RT-PCR: Interleukin 25, Junction plakoglobin and U1 small nuclear ribonucleoprotein

1C. The Snrp1c chimera was also examined via primer extension, since it appeared that the retroelement may be a candidate for an alternate start site for the gene.

5.2.1 Primer Extension

The primer extension reactions were run as described by Triezenberg (1992). The RNA for the reactions was taken from congenic A^{vy} mice, either a/a black, A^{vy}/a mottled, or A^{vy}/a yellow, depending upon which type of mouse the retroelement chimera appeared to be expressed in (in nearly all the cases the expression patterns were the same in the three types of mice, so it did not matter which RNA sample was used). Tissue choice was random and was driven by the fact that the liver and kidneys could provide a large amount of RNA for the experiments. Reactions were run a minimum of two times, and reactions which produced a positive band were run a minimum of four times. Therefore the results were reproducible and reliable.

Primers for the spliced transcripts Anxa6, Hao1, Sdh1 and Snrp1c, were designed to anneal to sequence on either side of the splice site to ensure the correct transcript was assayed. See Figure 5-2 for a pictorial representation. Where possible other primers were designed within 100bp of the candidate transcription start site, but not all sequences permitted this, so some were designed farther away. See Table 2-8 for primer sequences.

31 chimeric candidates were examined by primer extension reactions along with Snrp1c and the β actin control. Only 4 of the 32 candidates (Anxa6, Pik3ap1, Exoc4 and Snrp1c) gave bands which were reproducible and of a size to suggest the retroelement promoter was active. The Nr6a1 primer gave a band in half the reactions, but this inconsistency suggests either that it is expressed in varying amounts (different mouse RNA samples were used for the primer extension reactions), or that the primer was poor. A band was also produced in the Cdk7 reaction, but upon further analysis it

was found that the primer had been designed in an ID4 SINE element, so a new primer was designed. When the primer extension reaction was run with the new primer, no band was produced, suggesting that the initial band may have been an artefact. Refer to Figure 5-3 for primer extension results.

Since the production of bands in primer extension reactions depends upon the amount of transcript present in the total RNA mix, different primers produce bands of differing intensities. The Pik3ap1 band is comparable in intensity to the β actin control, which suggests that the Pik3ap1 chimeric transcript is being transcribed at quite a high level. On the other hand the Exoc4 and Snrp1c bands are faint, suggesting that the chimeric transcript is only transcribed at low levels. These bands were reproducible and were often darker when the reactions were repeated, indicating they are authentic promoter regions. The additional bands in Figure 5-3 are examples of weak stops; while some of these bands are stronger than the bands of interest (the predicted start site in the retroelement promoter), this does not necessarily indicate that the retroelement chimera is not authentic. Since it is likely that the retroelement chimeras are expressed at very low levels compared to genes, background expression from other "aberrant" promoters and retroelement promoters might produce bands of similar intensity.

Unfortunately, since some of the transcripts are present in low amounts, the film in Figure 5-3 had to be exposed to the radiation for a week so there is some diffusion of bands. This has made the Exoc4 and Snrp1c bands somewhat difficult to visualise as they have diffused into the background noise. When the film is developed after 2 days, there are very faint but distinct bands for Exoc4 and Snrp1c (results not shown).

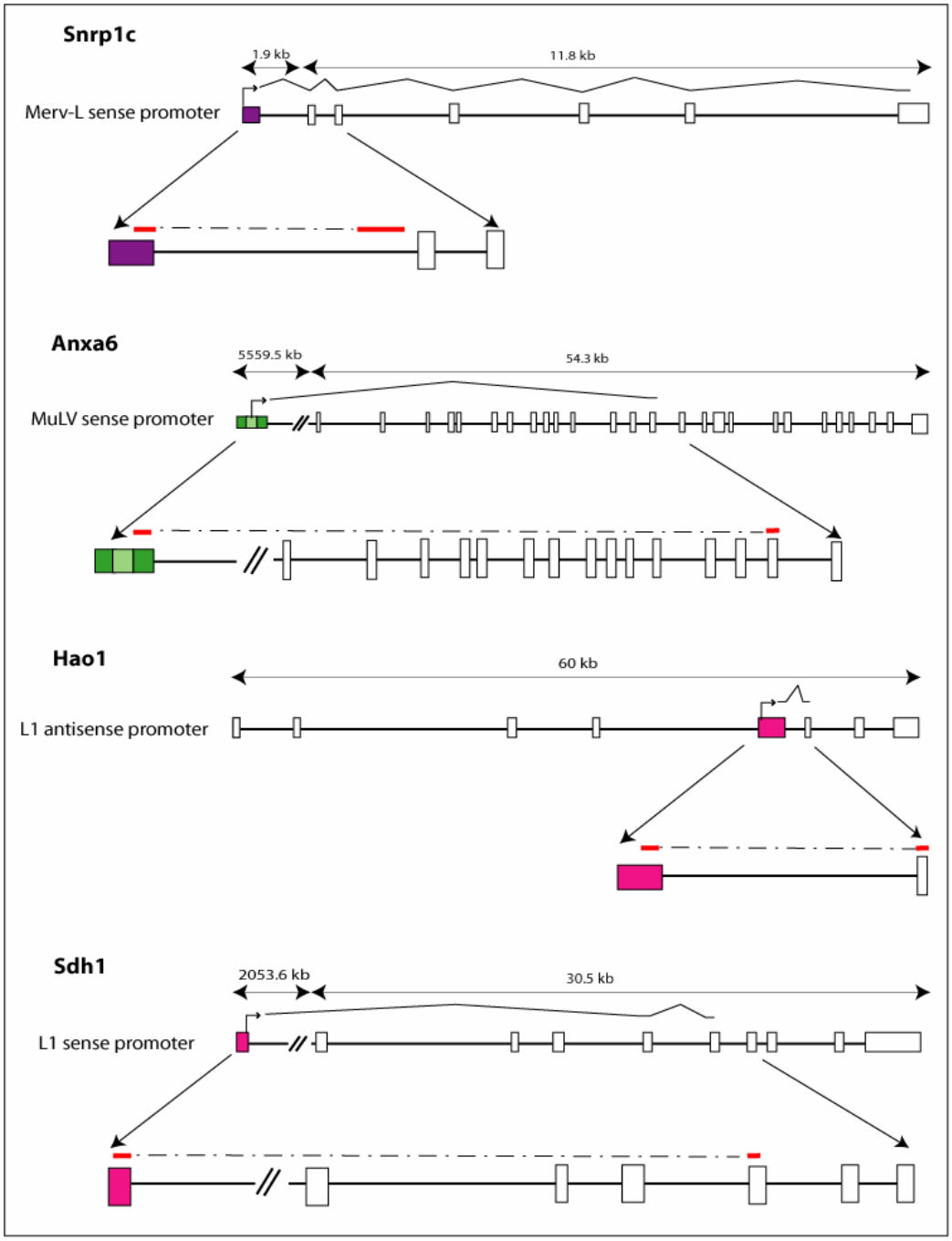

Figure 5-2 Schematic representation of the Snrp1c, Anxa6, Hao1 and Sdh1 gene loci showing primer extension primers. Exons are shown as white boxes, the Merv-L LTR is shown as a purple box, the MuLV element is shown as green boxes, the L1 element is shown as pink boxes and the primer extension primer is shown as red lines. The dashed line shows the sequence which was omitted from the primer.

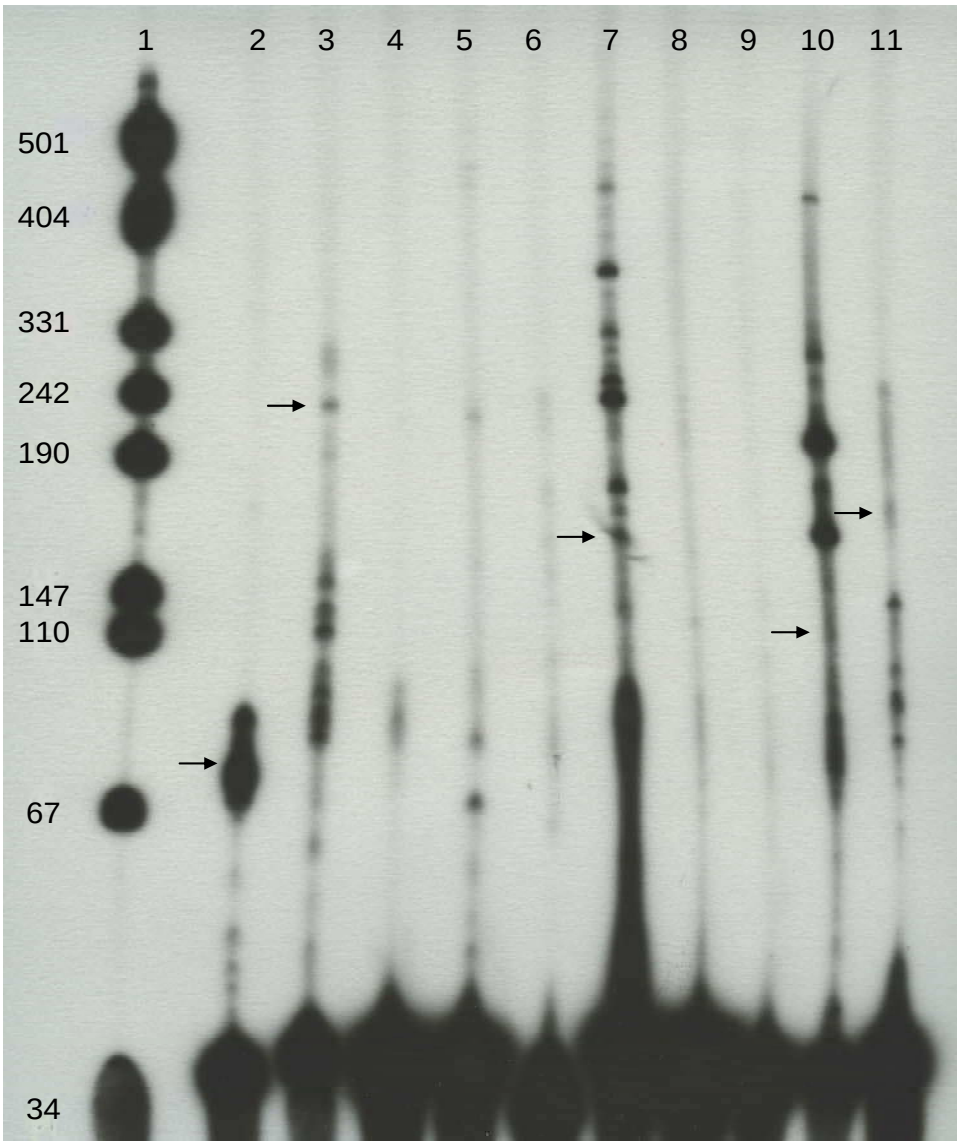

Figure 5-3 Primer extension results for chimeric mouse transcript promoter start sites. Lane 1 pUC 19 DNA/MspI marker, Lane 2 - β actin control, lane 3 - Anxa6, lane 4 - Cdk7, lane 5 - Hao1, lane 6 - Nr6a1, lane 7 - Pik3ap, lane 8 - Nipa2, lane 9 - Sdh1, lane 10 – Exoc4 and lane 11 - Snrp1c. Arrows indicate a predicted start site of transcription. The darker bands (β actin, Anxa6 and Pik3ap1) indicate high levels of transcript template and faint bands (Exoc4 and Snrp1c) indicate very low levels of transcript template. The β actin control is of a similar intensity to the Pik3ap band suggesting the chimera is transcribed at quite a high level. Other bands are weak stops. The film was exposed to the radioactivity for a week at -80ºC.

5.2.2 RT-PCR

The transposon displays identified a number of chimeric transcripts in a/a, A^{vy}/a mottled, and A^{vy}/a yellow mice. To our interest, most of the chimeric transcripts appeared to be expressed in the three different mice. This raised the questions - **are these chimeras strain specific or do other mouse strains also express the same chimeras? And does the expression of the chimeras vary between mice of the same strain (as is seen with the A^{vy} and $Axin^{Fu}$ mice)?** To answer these questions we used RT-PCR to look at expression of a random subset of chimeric candidates in Balb/c, C57BL/6J (a different subspecies to the mice used in the transposon displays), DBA, Fvb, Q/s and Sv129 mice. In keeping with previous experiments, RNA for the experiments were taken from the liver of the mice and all experiments were run on the same tissue sample. The candidate chimeras we examined were the Anxa6, Car5a, Cdk7, Cpne8, Nr6a1 and Pik3ap1 chimeras. In addition the IL25, Jup and Snrp1c chimeras, which were identified in the computer analysis, were also examined. Refer to Figure 5-4 for locus maps.

RT-PCRs were run according to the protocol supplied with the Invitrogen SuperScript One-Step RT-PCR with Platinum *Taq* kit, with 50 ng of mRNA being used in each reaction. GAPDH was used as a positive control and all reactions were set up in duplicate so a negative control (the reaction minus the RT enzyme) could be run. An additional nested PCR step, using a different forward primer, was conducted on the Cpne8, Nr6a1 and Snrp1c samples, as more than one band was amplified in the initial reaction. Refer to Figure 5-5 for RT-PCR results and Figure 5-4 for chimeric transcript loci maps.

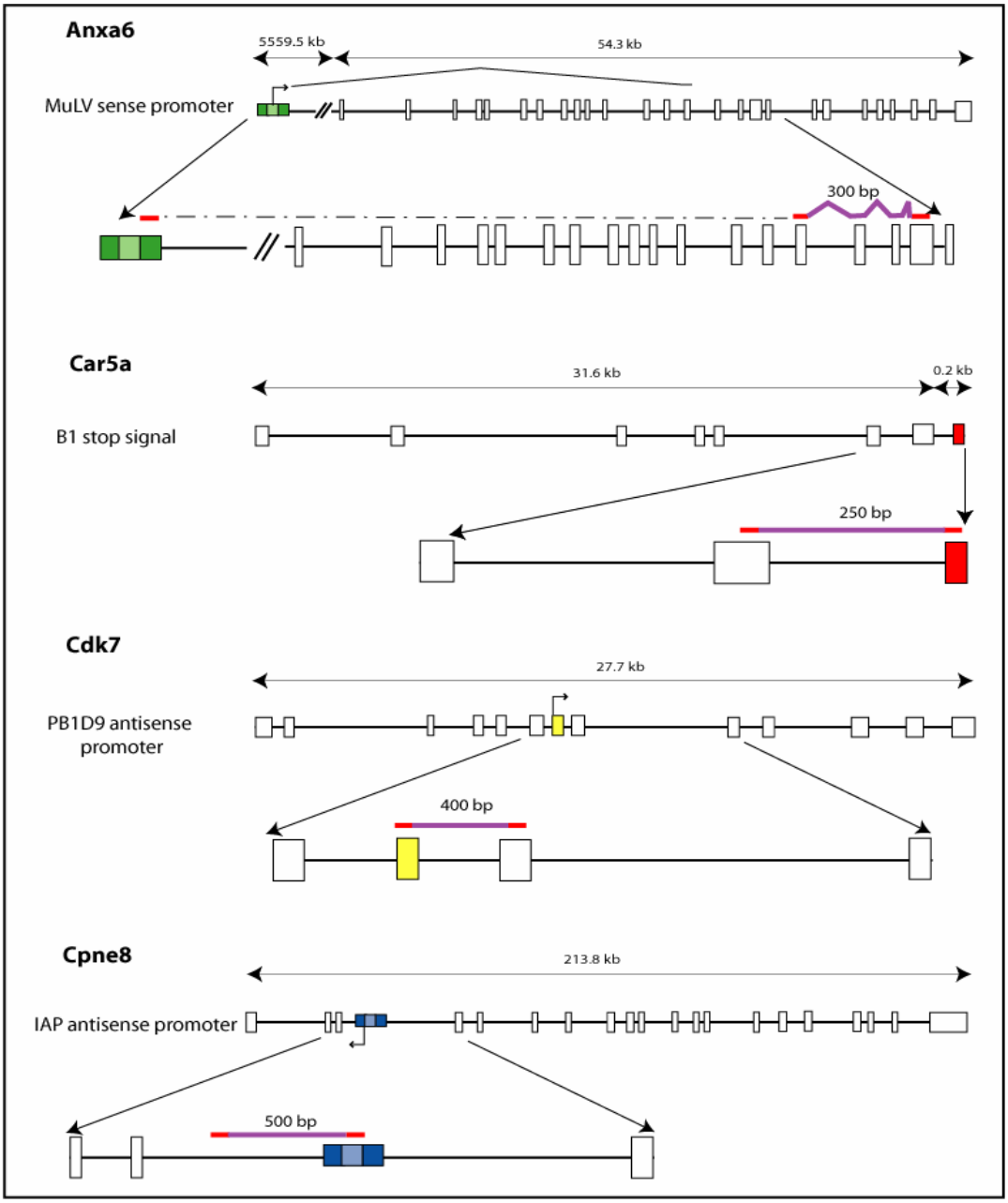

Figure 5-4 Gene locus maps for Anxa6, Car5a, Cdk7, Cpne8, Pik3ap1, Nr6a1, IL25, Jup and Snrp1. For each example the region being amplified by RT-PCR is shown in detail underneath the complete gene locus. Red lines indicate primers, purple lines indicate amplified regions and the dashed lines indicate sequence omitted from the primer. Red boxes are B1 elements, the orange box is a B2 element, the green box is a MuLV element, the purple box is a Merv-L LTR and the yellow box is a PB1D9 element. White boxes are gene exons.

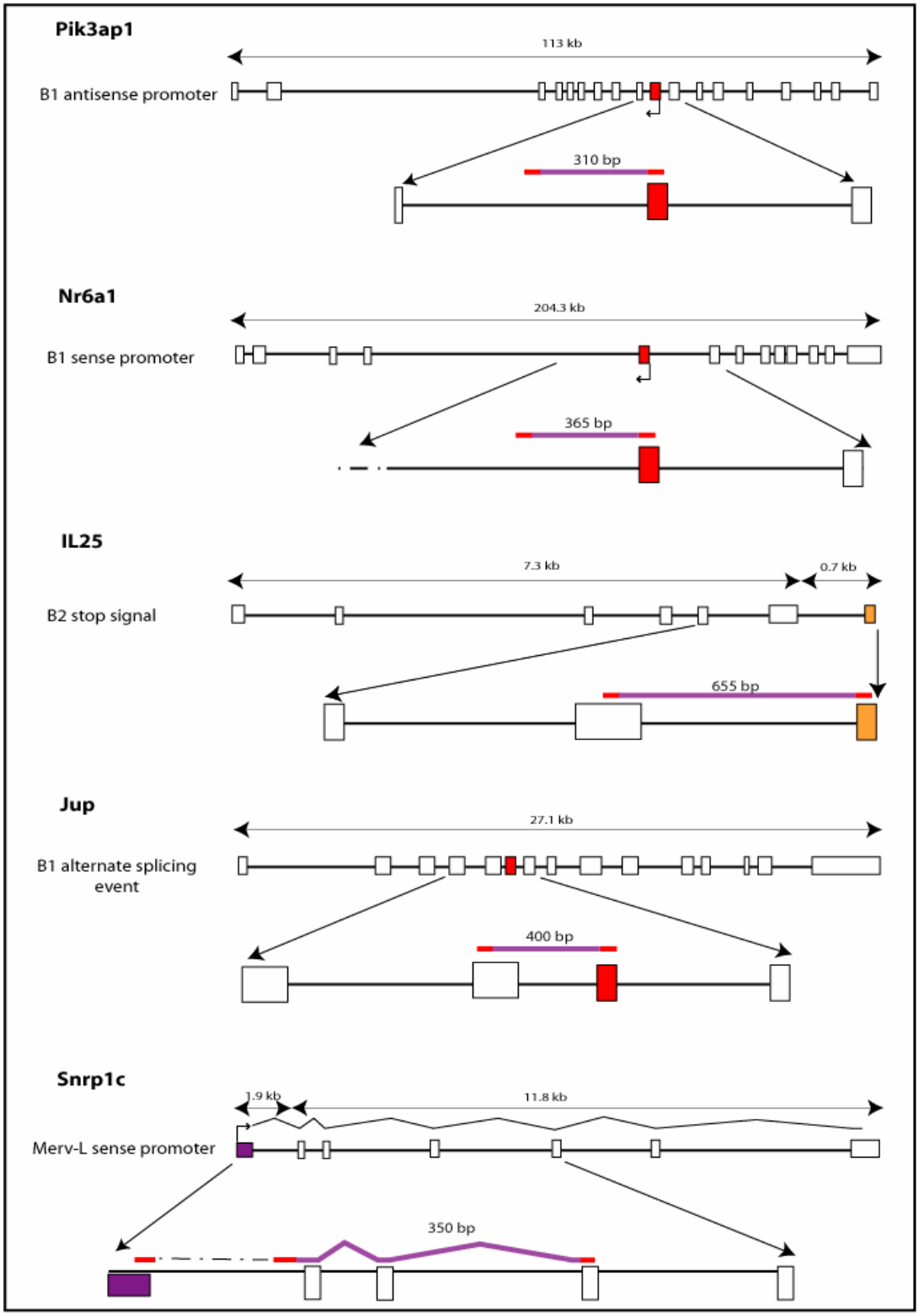

Figure 5-4 continued.

All of the chimeric transcripts we assayed were expressed in the 6 different mouse strains, except for Car5a which was not expressed in the Balb/c mice. The results are not quantitative, so conclusions based on differences between band intensities cannot be drawn; further quantitative RT-PCRs would need to be conducted to observe if there are expression level differences between the mice assayed.

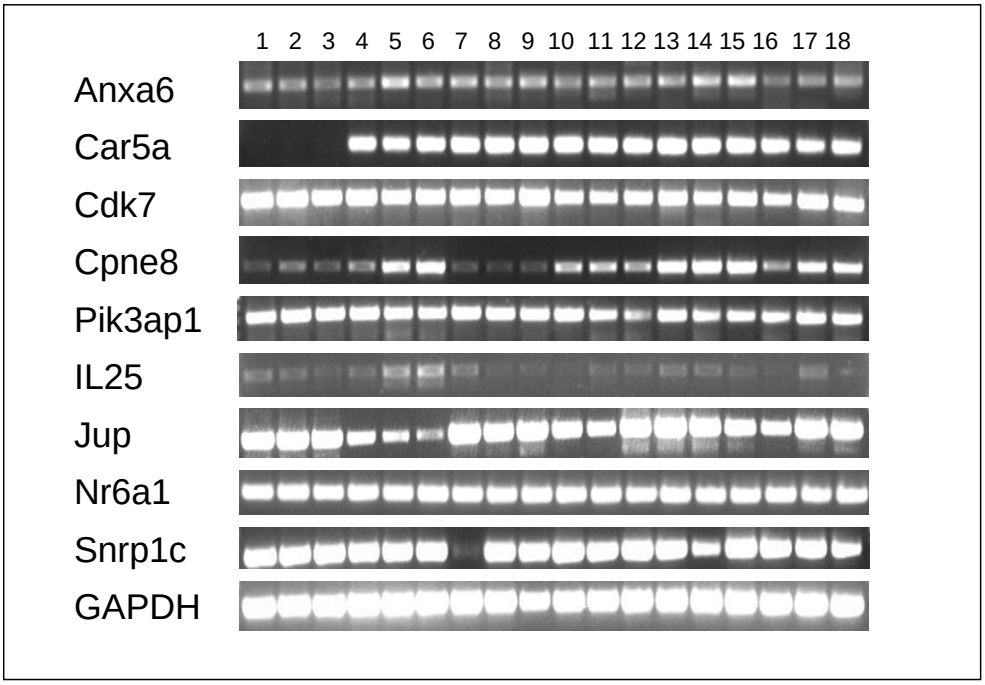

Figure 5-5 Non quantitative RT-PCR of a subset of chimeric retroelement candidates. RT-PCRs were conducted on chimeric transcripts in Balb/c (lanes 1-3), C57BL6/J (lanes 4-6), DBA (lanes 7-9), Fvb (lanes 10-12), Q/s (lanes 13-15) and SV/129 (lanes 16-18) mouse livers. Primers were designed to retroelement sequence and where possible exonic sequence for the genes Anxa6, Car5a, Cdk7, Cpne8, IL25, Jup, Pik3ap1, Nr6a1 and Snrp1c. GAPDH was used as a positive control. –RT controls are not shown.

The RT-PCR experiments were run a number of times to check reproducibility, and during the first run for the Cpne8 and Anxa6 primers, there appeared to be differential expression of the transcripts with bands missing in some mRNA samples. Upon rerunning the experiment with a new mRNA sample (prepared from the same

total RNA), it appeared that the transcripts were expressed in all the mice assayed. This suggests that some of the chimeric transcripts are likely to be present in low amounts, which results in what appears to be differential expression, when in fact the assay has just not amplified the transcript. This was overcome by adding more template and adding 5 cycles to the PCR run.

5.3 Discussion

The transposon displays produced a number of chimeric transcript candidates, which we analysed further by running primer extension and RT-PCR reactions, to provide evidence that at least some of the candidates were authentic transcripts.

Upon initial inspection, results from the primer extension assays suggested that not all the chimeric candidates were authentic. The Anxa6, Pik3ap1, Exoc4 and Snrp1c reactions produced bands of a size suggesting the chimeric transcripts were initiated in the promoters of retroelements. **It is difficult to predict what effect transcription of the retroelement chimeras would have upon the genes with which they are chimeric, because we know little about the functions of these genes**.

Annexin A6 is thought to be involved in calcium ion binding and transport, and it has been extensively examined in the heart (Kaetzel and Dedman 2004), where it is thought to be a regulator of intracellular calcium homeostasis. It is also thought that cell surface Anxa6 may function as an acidic pH binding site for Low Density Lipoprotein Receptor-related Protein-1 ligands and other proteins (Ling *et al.* 2004). Hawkins *et al.* (1999) knocked out the annexin 6 gene in mice and observed no phenotypic difference between the knockout mice and wildtype littermates (Hawkins *et al.* 1999). Therefore a change in regulation of Anxa6 due to the chimera would be difficult to observe, as it is unlikely to produce any obvious phenotypic change.

Phosphoinositide-3-kinase adaptor protein 1 plays an immunoregulatory role in B cell development and humoral immune responses (Yamazaki *et al.* 2002). Mutant mice deficient in the protein were found to be viable, but they were B1 B cell deficient and had decreased numbers of mature B cells. If the retroelement chimera is interfering with the expression of Pik3ap1, the B cell population may be affected; and so this gene may be a candidate for future investigation to ask if transcriptional interference is

induced by the B1 retroelement. However when the RT-PCR results are examined, it appears that the Pik3ap1 gene is not differentially expressed between mouse species which would make it difficult to analyse the expression of the chimera, especially as this observation may indicate that transcription of the Pik3ap1 chimera is normal. Note however that the RT-PCR is not quantitative and there may be some hidden variation.

Exocyst complex component 4 is part of the exocyst complex, which selectively regulates the docking of insulin-containing vesicles at sites of release close to the plasma membrane (Tsuboi *et al.* 2005). It is also thought to play a role in the delivery of N-methyl-D-aspartate receptors to the cell surface in heterologous cells and neurons, via interactions with synapse-associated protein 102 (Sans *et al.* 2003). A knock out of the Exoc4 gene is lethal in homozygotes, who display a delay in gastrulation and fail to progress past the primitive streak stage (Friedrich *et al.* 1997). Heterozygotes however displayed no phenotype, so again it would be difficult to study the affects that transcription of the chimera may have on regulation of Exoc4.

The U1 small nuclear ribonucleoprotein 1C is involved in pre-mRNA splicing, where it defines the intron/exon boundaries by binding to consensus sequences within the pre-mRNA (Krainer and Maniatis 1985). No knockout mice have been made, so it is not known whether transcriptional interference by the Snrp1c chimera would produce an obvious phenotype. The band produced in the primer extension is very weak, again suggesting that this might not be a good candidate to analyse further.

A large proportion (5/6) of the primer extension reactions did not produce obvious “strong stops”. As mentioned previously, the lack of bands may be due to low levels of chimeric transcript being present in the total RNA pool. This translates to a small amount of template material for the primer extension reaction, which in turn results in no band being formed. The transposon display involves subjecting the initial

chimeric transcript to a number of rounds of PCR amplification, which would help explain why a band is visualised for the display but not for the primer extension reaction (where there is no amplification step).

Alternatively no band may have been visualised because the transcript may not be stable, and this would also result in no band being amplified. No band would also be expected if the retroelement is acting as a stop site. Two of the chimeric retroelements assayed, Abcd4 and Smp1, were known to be acting as stop sites - indicated by splicing patterns (in the past the gene had appropriated the retroelement and it is now part of gene) and another, Car5a, appeared to be acting as an alternate stop site. However we were still interested in seeing if the elements were capable of producing a transcript, since it is likely that the elements would be unmethylated whilst they are acting as a stop site, and therefore may also be capable of transcription. This type of chimeric transcript, while not being an example of a retroelement-driven transcription, is still a possible candidate for transcriptional interference. Genes which are prematurely terminated, such as Car5a, may produce aberrant transcripts which may in turn be translated into an aberrant protein. Even if the transcript is not translated, the normal expression patterns of the gene are being interrupted by the retroelement.

Other primer extension reactions <u>which might be expected to give negative results</u> were those that looked at elements in the 3' untranslated region of genes such as the Siat4c and Rrn3 reactions. Its expected that it would be difficult to determine if the retroelements were being expressed, due to the high background transcription of the gene with which the retroelement was associated. But since the chromatin conformation is open during transcription of the gene, the 3' UTR containing the retroelement could also become active, so we decided to run the experiments in the hope of seeing transcription being initiated from the retroelement promoter. None of

the reactions produced a band, although only a few candidates were analysed, so the possibility that retroelements in 3' UTRs may have active promoters is not excluded. While the consequences of this type of transcription are not known we speculate that, as with the elements in 5' UTRs (Landry *et al.* 2001), the retroelements in the 3' UTRs may have a regulatory function.

While the primer extension reactions can give a band of an expected size to suggest that the retroelement promoter is active, the band cannot be sequenced to provide further evidence that the promoter of the retroelement of interest actually produced the transcript. We attempted 5' RACE experiments, which would be able to give a more precise answer, but we had difficulty obtaining any results and abandoned the experiments due to time constraints.

The RT-PCR experiments were quite successful, with all the chimeras analysed being expressed in at least five of the six mouse strains examined. Of the 9 chimeras analysed, only one displayed differential expression. This suggests that it is likely that most of the chimeras we analysed were present in the ancestral mouse species, and that their expression has been conserved throughout the formation of the different mouse strains. This seems a likely explanation, since the retroelements which are involved in chimeric transcript formation are likely to be present in areas of active chromatin; otherwise it would be expected that methylation would keep the vast majority of the retroelements silent (Yoder *et al.* 1997; Bird 2002). Since most of the genes which would be active in one mouse strain would also be active in all other mouse strains, it would mean the same subset of retroelements are subject to activation. As mentioned earlier, there is also likely to be a subset that are randomly activated, and these are candidates for producing phenotypic variation and in some cases disease (Whitelaw and Martin 2001). Unfortunately the small number of candidates we managed to study in

detail did not reveal any candidates from this latter group, but it does not mean that they do not exist. It should also be noted that quantitative RT-PCR was not conducted on the candidates so there may be some individual variation in expression between mice which was missed, by using normal RT-PCR, but which could contribute to phenotypic differences.

While authentication of retroelement transcription via the primer extension reactions gave a small subset of active retroelement promoters, the RT-PCR results suggest that expression of the retroelement chimeras may be widespread amongst different mouse strains. Of the nine retroelement chimeras which gave positive results in most mouse strains for RT-PCR, two chimeras did not have primer extension reactions run on them, the Car5a example was thought to be an alternate stop site so would be unlikely to produce a primer extension band, leaving only three of the six remaining retroelement chimeras giving a positive result in both the primer extension and RT-PCR reactions. This discrepancy may be explained in a number of different ways. Perhaps some of the chimeras are being initiated in a different promoter – perhaps even a different retroelement promoter, which is 5' to the promoter of the retroelement being examined. Another possibility is that some of the chimeras are artefacts; examples of retroelements which are present in introns that have not been spliced out of the transcript before the poly A tail was added. Some examples are likely to be alternate stop sites; which would be amplified in an RT-PCR experiment but would not produce a band in the primer extension reactions, while other chimeras may be formed due to expression of retroelements in untranslated exons. As mentioned above the primer extension reactions may not have amplified all the authentic transcripts which would also result in this discrepancy.

We were unsure whether the chimeric sequences taken from the dbEST would give any results in the primer extension and RT-PCR reactions, because they may have been randomly activated and expressed only in the particular individual from which they were recovered. The IL25 and Snrp1c chimeric transcripts were found in a neonate retina library and a library containing a variety of embryos respectively, and since the mice we examined were over 3 weeks of age, we thought it likely that they would not be expressed in our samples. The Jup chimeric transcript was from a library made from skin of 11 week females so it seemed possible that it would be expressed in our samples, although we were looking in liver and kidney rather than skin.

To our surprise the 3 chimeras were amplified in the RT-PCR, and the Snrp1c chimera appeared to give a faint band in the primer extension reaction. This suggests that the chimeras are commonly expressed, and also that they are not examples of chimeras which are likely to cause phenotypic differences. The apparent low expression of Snrp1c may explain why additional transcripts of Snrp1c were not detected in other EST libraries, as would be expected for commonly expressed chimeras. If the chimera is expressed at a low enough level it may be missed when EST libraries are made.

The **results of these experiments illustrate the dilemma created by the experimental strategy**: it recovers so many candidates that further study requires selection of a few, which in turn requires us to guess which ones might be most significant. Given this, we think it is not surprising that we have yet to uncover a clear case in which a retroelement transcript varies between individual mice to create some functional (phenotypic) difference. Because such a large number of chimeric candidates were found using the transposon display, it was difficult for me to choose which transcripts should be examined further. No method is available which would allow me

to authenticate all, or even most of the recovered transcripts in a single experiment. Primer extension, RT-PCR, Northern Blots and 5' RACE all require specific primers or probes, which means no fast, large scale experiment could be conducted. In addition the similarity between the retroelement portions of the chimeras would hamper any large scale experiment. This problem meant only a small number of chimeric candidates were examined, and therefore my results may not be a clear indication of how many retroelement chimeras are in fact authentic. A large amount of further work would be necessary to determine what proportion of total chimeric transcripts were authentic, but we have provided a number of candidates which could be examined in further detail.

In this chapter we have shown with primer extension and RT-PCR that only a small proportion of the transcripts recovered from the transposon display are likely to be authentic retroelement chimeras. For reasons discussed above, the proportion of authenticate transcripts may be higher, but due to limitations of the methods we used for authentication some transcripts may have been missed. In addition, only a small number of the chimeras obtained from the transposon displays were used for the primer extension and RT-PCR reactions, so there are likely to be many more legitimate chimeric transcripts in the pool we amplified using the transposon display method. Further work, including running Northern blots and 5' RACE experiments, is needed to test how effective the transposon display method is for locating chimeric transcripts in the mouse. Nevertheless it seems likely that there are a great many possibly aberrant transcripts initiated in retroelement promoters. My results are only the “tip of the iceberg”, and if one were able to study all of the transcripts it is quite possible that others that have clear biological significance would be found.

CHAPTER 6 CONCLUSIONS

Summary and concluding remarks

This project aimed to gain a better understanding of the activity of retroelements in the mammalian genome, and to seek evidence in support of the idea that they are capable of transcriptional interference in somatic cells. The finding that retroelements regularly perturb gene expression would indicate that they are more than just pieces of "junk" DNA, i.e. they may not only have biological significance but may even be capable of causing disease.

In silico analysis

While a number of studies had shown a range of retroelements to be active in normal and diseased tissue, no large-scale study had examined retroelement transcription to see how common it is for retroelements to have intact, active promoters. We set out to determine how many retroelement families have regulatory regions capable of initiating or terminating transcription. We then decided to extend the study to ask how many unique retroelement regulatory regions are active. Since it is not feasible to assay every retroelement in a human or mouse, we approached the question by examining the dbEST for retroelement transcripts initiated or terminated within a regulatory region. Because the human genome sequence was completed before the mouse genome sequence, we conducted a thorough analysis on human retroelements and were able to complete a smaller scale analysis of mouse retroelements.

In Chapter 3 of this thesis, we have shown that most retroelement families from the human and mouse produce transcripts which appear to be initiated or terminated in promoters of the retroelements, transcribed in a sense or antisense orientation, and present in normal as well as diseased tissue. These findings refute the notion that retroelements are maintained in a silent state in somatic cells. It was also significant that most retroelement families have at least some members that are capable of

transcription, as it is often assumed that only a few families are still active. The number of promoters that were initially thought to be capable of becoming active is enlarged further if the possible antisense SINE promoter is taken into account. Possibly the most important finding from the *in silico* analysis was that many retroelement families produce transcripts that are chimeric with cellular genes. This was direct evidence of transcriptional interference, and taken together with the above results indicates the phenomenon may be fairly common.

Transposon display

After finding chimeric transcripts in dbEST, we sought direct evidence for chimeric retroelement-gene transcript formation *in vivo*. Few studies (none in mammals) have taken a broad approach to this question, so we used a transposon display to locate retroelement chimeric candidates in mice. Transposon display amplifies transcripts that contain retroelement sequences. This method allowed us to study the repertoire of retroelement transcription in individual mice, and to compare mice with each other. We recovered very large numbers of transcripts; this illustrates one problem with the strategy: it produces more candidates than one can investigate in detail, and choosing which ones to investigate further involves guesswork. Of the small numbers of chimeras we examined, some candidates were chimeric with intergenic sequences, some were chimeric with intronic sequences, and a number of candidates were chimeric with cellular genes. While all candidates were equally likely to be exerting transcriptional interference on nearby genes, we focused on the candidates which were chimeric with genes, as these examples were the most likely examples of transcriptional interference.

Primer extension and RT-PCR

Further experiments, described in Chapter 5, used primer extension and RT-PCR to provide evidence that the chimeric transcripts were authentic. While only a small subset passed this test, the results support our hypothesis that retroelement-based transcriptional interference may be fairly common (it should be remembered that the candidates chosen for further study were only a small proportion of the total). In fact most mice appear to have a common pattern of retroelement chimeric expression, with differences between strains being minor. This does not discount our idea that individuals have different subsets of active retroelements being expressed in mosaic patterns, as these types of retroelements may be expressed as only a small percentage of the whole, and so may have been missed in our assay. Only a few cases of transcriptional interference have been described in the literature, and my results are significant because they should help expand the number of elements which will be candidates for future studies in depth.

Since little was known about the retroelement chimeras which appeared to be authentic, we could not draw conclusions as to whether the retroelements are having a beneficial, neutral or detrimental affect on the hosts. There are numerous studies supporting the three views, and it seems likely that retroelements exerting transcriptional interference may fall into all three categories.

Because there are so many retroelements in the genome, it would be surprising if some of the chimeras we uncovered are not simply the products of "background transcription" (transcription that is initiated in various cryptic promoters and does not produce functional transcripts). Johnson *et al.* (2004) found that there is evidence of widespread transcription outside the boundaries of known genes, with transcriptional activity being observed even in intergenic regions (Johnson *et al.* 2005). Unfortunately

their study does not include retroelements (a common theme in studies of gene expression), but it may explain at least some transcription of retroelements in intergenic regions. This would suggest that retroelement chimeras are just "noise" and have no function. But even if many chimeric transcripts are the products of background expression, if their production interferes (in any of a number of ways) with transcription of the genes around them then they may be regarded as having a biological function - although not necessarily an adaptive one (Whitelaw and Martin 2001).

With the recent evidence that polymorphisms play a smaller role than expected in differences between individuals, there have been more studies concluding that retroelements may play a role in cell biology, as described in Section 1.8. It is likely that some of the chimeras we discovered are acting in a beneficial manner to globally regulate genes in a similar way to that described by Peaston *et al*. (2004), while others may be acting to downregulate genes via RNAi mechanisms described by Svoboda *et al.* (2004). Yet others may be influencing genes in a detrimental way, such as seen with the A^{vy} and $Axin^{Fu}$ mice, but in a manner that does not produce an obvious phenotype. Further work will be necessary to elucidate the role of the retroelement transcripts in the mouse; this might involve Northern blots to assay expression, 5' and 3' RACE to clearly show where transcripts begin and end, and biological assays (which would have to be tailored to each gene being studied).

My study has established that retroelement-gene chimeras are widely expressed in somatic cells. Because transcription of retroelements is under epigenetic control, it is likely that it can be perturbed either to increase or decrease activity. This might occur as part of natural variation in environmental or nutritional conditions. Much more work will be necessary if we are to know how much retroelement transcription can vary, and how it might be manipulated. But if retroelement activity is primarily detrimental to

normal programs of gene expression, then artificial measures to suppress it could have health benefits.

This thesis project produced an enormous amount of data from both the *in silico* analysis and the transposon displays. The volume of candidates produced by the two strategies made it difficult to determine which retroelements should (or could) be studied in more detail. Druker *et al.* (2004) examined the IAP/CDK5 activator binding protein (Cabp) chimera in detail (it was one of the chimeras pulled up in our transposon display), and found the transcript to be variably expressed in isogenic littermates. Bisulfite sequencing, 5' and 3' RNA ligase-mediated RACE, and Northern blots were conducted on the chimera to further show that when the IAP element is unmethylated, a number of short transcripts are initiated from the promoter of the Cabp gene which terminate prematurely 5' of the IAP. This is an example off the sort of further investigation that could be conducted on the transcripts we have identified. It illustrates the dilemma created by even the limited search we conducted: it produces far more candidates than can be studied in detail, but it does not provide any way of choosing which of them might be studied further most productively.

We could have conducted experiments similar to those done by Druker *et al.* (2004) on a number of different retroelement chimeric candidates, taken from either the *in silico* analysis or the transposon display results, but we decided to examine our results as a whole in order to gain a better grasp of how many retroelements could be interfering with gene expression. As the results should make clear, the number is potentially very large indeed. It would be possible to go back to our data and pick chimeras to analyse more fully. Further work will be conducted to examine candidates in more detail, and it is likely we will uncover many more examples like the Cabp chimera.

Further Work

This thesis created a large amount of data which could not be thoroughly analysed in the timeframe of a PhD, with further work being needed in all the areas we examined.

In the *in silico* analysis further work could be conducted to not only show how many chimeric transcripts are present in the dbEST, but to also give detailed statistics on libraries the transcripts are located in as well as statistics for chromosomal locations for both mouse and human retroelements.

Further work on the transposon display could be conducted with experiments being run in different tissues to see if new retroelement chimeras could be identified. Additional bands could also be excised and sequenced to locate new chimeric candidates.

The primer extension and RT-PCR results could be verified using RACE and quantitative PCR experiments and additional candidates could be examined to determine which other chimeric candidates are also genuine.

Therefore this PhD has acted as the groundwork for future work which will be able to give detailed information about the activity of retroelements in humans and mice.

In conclusion, we have shown that many different families of retroelements have intact regulatory regions, which are capable of active transcription; consequently retroelements have the capability to influence transcription of surrounding genes in somatic cells. Through its effects on gene expression, retroelement transcription could have a large impact on the phenotype of an organism. Further research may reveal the magnitude of this impact, and its role in phenotypic variation.

REFERENCES

Ali, M., Veale, D. J., Reece, R. J., Quinn, M., Henshaw, K., Zanders, E. D., Markham, A. F., Emery, P. and Isaacs, J. D. (2003). "Overexpression of transcripts containing LINE-1 in the synovia of patients with rheumatoid arthritis." Ann Rheum Dis **62**(7): 663-666.

Andersson, A. C., Svensson, A. C., Rolny, C., Andersson, G. and Larsson, E. (1998). "Expression of human endogenous retrovirus ERV3 (HERV-R) mRNA in normal and neoplastic tissues." Int J Oncol **12**(2): 309-13.

Andersson, G., Svensson, A.-C., Setterblad, N. and Rask, L. (1998). "Retroelements in the human MHC class II region." Trends in Genetics **14**(3): 109-114.

Argeson, A. C., Nelson, K. K. and Siracusa, L. D. (1996). "Molecular basis of the pleiotropic phenotype of mice carrying the hypervariable yellow (Ahvy) mutation at the agouti locus." Genetics **142**(2): 557-67.

Ashe, H. L., Monks, J., Wijgerde, M., Fraser, P. and Proudfoot, N. J. (1997). "Intergenic transcription and transinduction of the human beta -globin locus." Genes Dev. **11**(19): 2494-2509.

Athanikar, J. N., Badge, R. M. and Moran, J. V. (2004). "A YY1-binding site is required for accurate human LINE-1 transcription initiation." Nucl. Acids Res. **32**(13): 3846-3855.

Badenhoop, K., Tonjes, R. R., Rau, H., Donner, H., Rieker, W., Braun, J., Herwig, J., Mytilineos, J., Kurth, R. and Usadel, K. H. (1996). "Endogenous retroviral long terminal repeats of the HLA-DQ region are associated with susceptibility to insulin-dependent diabetes mellitus." Hum Immunol **50**(2): 103-10.

Barbot, W., Dupressoir, A., Lazar, V. and Heidmann, T. (2002). "Epigenetic regulation of an IAP retrotransposon in the aging mouse: progressive demethylation and de-silencing of the element by its repetitive induction." Nucl. Acids Res. **30**(11): 2365-2373.

Baust, C., Seifarth, W., Germaier, H., Hehlmann, R. and Leib-Mosch, C. (2000). "HERV-K-T47D-Related Long Terminal Repeats Mediate Polyadenylation of Cellular Transcripts." Genomics **66**(1): 98-103.

Beck, J. A., Lloyd, S., Hafezparast, M., Lennon-Pierce, M., Eppig, J. T., Festing, M. F. and Fisher, E. M. (2000). "Genealogies of mouse inbred strains." Nat Genet **24**(1): 23-5.

Bestor, T. H. (2000). "The DNA methyltransferases of mammals." Hum. Mol. Genet. **9**(16): 2395-2402.

Bieda, K., Pani, M. A., van der Auwera, B., Seidl, C., Tonjes, R. R., Gorus, F., Usadel, K. H. and Badenhoop, K. (2002). "A retroviral long terminal repeat adjacent to the HLA DQB1 gene (DQ-LTR13) modifies Type I diabetes susceptibility on high risk DQ haplotypes." Diabetologia **45**(3): 443-447.

Bing, O. H. L., Sirokman, G. and Humphries, D. E. (1998). "Hypothesis: Link Between Endogenous Retroviruses and Cardiovascular Disease." Journal of Molecular and Cellular Cardiology **30**(7): 1257-1262.

Bird, A. (2002). "DNA methylation patterns and epigenetic memory." Genes Dev **16**(1): 6-21.

Bock, M. and Stoye, J. P. (2000). "Endogenous retroviruses and the human germline." Curr Opin Genet Dev **10**(6): 651-5.

Boeke, J. D. and Stoye, J. P. (1997). Retrotransposons, endogenous retroviruses, and the evolution of retroelements. In *Retroviruses*. Varmus, H. E. Cold Spring Harbour, Cold Spring Harbour Laboratory Press**:** 348-352.

Branciforte, D. and Martin, S. L. (1994). "Developmental and cell type specificity of LINE-1 expression in mouse testis: implications for transposition." Mol Cell Biol **14**(4): 2584-92.

Bratthauer, G. L., Cardiff, R. D. and Fanning, T. G. (1994). "Expression of LINE-1 retrotransposons in human breast cancer." Cancer **73**(9): 2333-6.

Britten, R. J. (1996). "DNA sequence insertion and evolutionary variation in gene regulation." Proc Natl Acad Sci U S A **93**(18): 9374-7.

Britten, R. J. (1997). "Mobile elements inserted in the distant past have taken on important functions." Gene **205**(1-2): 177-82.

Britten, R. J. (2004). "Coding sequences of functioning human genes derived entirely from mobile element sequences." Proc Natl Acad Sci U S A **101**(48): 16825-30.

Bromham, L., Clark, F. and McKee, J. J. (2001). "Discovery of a Novel Murine Type C Retrovirus by Data Mining." J. Virol. **75**(6): 3053-3057.

Brulet, P., Kaghad, M., Xu, Y.-S., Croissant, O. and Jacob, F. (1983). "Early Differential Tissue Expression of Transposon-Like Repetitive DNA Sequences of the Mouse
10.1073/pnas.80.18.5641." PNAS **80**(18): 5641-5645.

Budelier, K. and Schozz, J. (1992). Purification and concentration of DNA from aqueous solutions. In *Current Protocols In Molecular Biology*. Struhl, K. United States of America, John Wiley and Sons Inc.**:** 2.1.1-2.1.3.

Burden, A. F., Manley, N. C., Clark, A. D., Gartler, S. M., Laird, C. D. and Hansen, R. S. (2005). "Hemimethylation and Non-CpG Methylation Levels in a Promoter Region of Human LINE-1 (L1) Repeated Elements." J. Biol. Chem. **280**(15): 14413-14419.

Burwinkel, B. and Kilimann, M. W. (1998). "Unequal homologous recombination between LINE-1 elements as a mutational mechanism in human genetic disease." Journal of Molecular Biology **277**(3): 513-517.

Buscher, K., Trefzer, U., Hofmann, M., Sterry, W., Kurth, R. and Denner, J. (2005). "Expression of Human Endogenous Retrovirus K in Melanomas and Melanoma Cell Lines." Cancer Res **65**(10): 4172-4180.

Casa, A. M., Brouwer, C., Nagel, A., Wang, L., Zhang, Q., Kresovich, S. and Wessler, S. R. (2000). "Inaugural Article: The MITE family Heartbreaker (Hbr): Molecular markers in maize." PNAS **97**(18): 10083-10089.

Chalitchagorn, K., Shuangshoti, S., Hourpai, N., Kongruttanachok, N., Tangkijvanich, P., Thong-ngam, D., Voravud, N., Sriuranpong, V. and Mutirangura, A. (2004). "Distinctive pattern of LINE-1 methylation level in normal tissues and the association with carcinogenesis." Oncogene **23**(54): 8841-6.

Chen, J., Sun, M., Hurst, L. D., Carmichael, G. G. and Rowley, J. D. (2005). "Genome-wide analysis of coordinate expression and evolution of human cis-encoded sense-antisense transcripts." Trends in Genetics **21**(6): 326-329.

Cheng, J., Kapranov, P., Drenkow, J., Dike, S., Brubaker, S., Patel, S., Long, J., Stern, D., Tammana, H., Helt, G. *et al.* (2005). "Transcriptional Maps of 10 Human Chromosomes at 5-Nucleotide Resolution." Science **308**(5725): 1149-1154.

Cherry, J. M. (1992). Compressions. In *Current Protocols In Molecular Biology*. Struhl, K. United States of America, John Wiley and Sons Inc.**:** 7.4.27-7.4.31.

Coffin, J. M. (1992). Structure and classification of retroviruses. In *The Retroviridae*. Levy, J. A. New York, Plenum Press. **1:** 19-49.

Coffin, J. M. (1995). Retrovirus variation and evolution. In *The DNA provirus: Howard Temin's scientific legacy*. Sugden, B. Washington, D.C., American Society for Microbiology**:** 221-244.

Corbin, V. and Maniatis, T. (1989). "Role of transcriptional interference in the Drosophila melanogaster Adh promoter switch." Nature **337**(6204): 279-282.

Craddock, N., O'Donovan, M. C. and Owen, M. J. (2006). "Genes for Schizophrenia and Bipolar Disorder? Implications for Psychiatric Nosology." Schizophr Bull **32**(1): 9-16.

Croft, J. A., Bridger, J. M., Boyle, S., Perry, P., Teague, P. and Bickmore, W. A. (1999). "Differences in the localization and morphology of chromosomes in the human nucleus." J Cell Biol **145**(6): 1119-31.

de Parseval, N., Lazar, V., Casella, J.-F., Benit, L. and Heidmann, T. (2003). "Survey of Human Genes of Retroviral Origin: Identification and Transcriptome of the Genes with Coding Capacity for Complete Envelope Proteins." J. Virol. **77**(19): 10414-10422.

Dean, R. G., Socher, S. H. and Cutler, R. G. (1985). "Dysdifferentiative nature of aging: age-dependent expression of mouse mammary tumor virus and casein genes in brain and liver tissues of the C57BL/6J mouse strain." Archives of Gerontology and Geriatrics **4**(1): 43-51.

Dean, W., Lucifero, D. and Santos, F. (2005). "DNA methylation in mammalian development and disease." Birth Defects Research Part C: Embryo Today: Reviews **75**(2): 98-111.

DeBerardinis, R. J., Goodier, J. L., Ostertag, E. M. and Kazazian, H. H., Jr. (1998). "Rapid amplification of a retrotransposon subfamily is evolving the mouse genome." Nat Genet **20**(3): 288-90.

DeBerardinis, R. J. and Kazazian, H. H., Jr. (1999). "Analysis of the promoter from an expanding mouse retrotransposon subfamily." Genomics **56**(3): 317-23.

Deininger, P. L. and Batzer, M. A. (1999). "Alu Repeats and Human Disease." Molecular Genetics and Metabolism **67**(3): 183-193.

Deininger, P. L. and Batzer, M. A. (2002). "Mammalian retroelements." Genome Res **12**(10): 1455-65.

Deininger, P. L., Moran, J. V., Batzer, M. A. and Kazazian, H. H., Jr. (2003). "Mobile elements and mammalian genome evolution." Curr Opin Genet Dev **13**(6): 651-8.

Depil, S., Roche, C., Dussart, P. and Prin, L. (2002). "Expression of a human endogenous retrovirus, HERV-K, in the blood cells of leukemia patients." Leukemia **16**(2): 254-9.

Dewannieux, M., Esnault, C. and Heidmann, T. (2003). "LINE-mediated retrotransposition of marked Alu sequences." Nat Genet **35**(1): 41-8.

Dias Neto, E., Garcia Correa, R., Verjovski-Almeida, S., Briones, M. R. S., Nagai, M. A., da Silva, W., Jr., Zago, M. A., Bordin, S., Costa, F. F., Goldman, G. H. *et al.* (2000). "Shotgun sequencing of the human transcriptome with ORF expressed sequence tags." PNAS **97**(7): 3491-3496.

Druker, R., Bruxner, T. J., Lehrbach, N. J. and Whitelaw, E. (2004). "Complex patterns of transcription at the insertion site of a retrotransposon in the mouse." Nucleic Acids Res **32**(19): 5800-8.

Duhl, D. M., Stevens, M. E., Vrieling, H., Saxon, P. J., Miller, M. W., Epstein, C. J. and Barsh, G. S. (1994). "Pleiotropic effects of the mouse lethal yellow (Ay) mutation explained by deletion of a maternally expressed gene and the simultaneous production of agouti fusion RNAs." Development **120**(6): 1695-708.

Dunn, C. A., Medstrand, P. and Mager, D. L. (2003). "An endogenous retroviral long terminal repeat is the dominant promoter for human beta1,3-galactosyltransferase 5 in the colon." Proc Natl Acad Sci U S A **100**(22): 12841-6.

Dupressoir, A., Puech, A. and Heidmann, T. (1995). "IAP retrotransposons in the mouse liver as reporters of ageing." Biochim Biophys Acta **1264**(3): 397-402.

Ejthadi, H. D., Martin, J. H., Junying, J., Roden, D. A., Lahiri, M., Warren, P., Murray, P. G. and Nelson, P. N. (2005). "A novel multiplex RT-PCR system detects human endogenous retrovirus-K in breast cancer." Archives of Virology **150**(1): 177-184.

Englander, E., Wolffe, A. and Howard, B. (1993). "Nucleosome interactions with a human Alu element. Transcriptional repression and effects of template methylation." J. Biol. Chem. **268**(26): 19565-19573.

Englander, E. W. and Howard, B. H. (1995). "Nucleosome Positioning by Human Alu Elements in Chromatin." J. Biol. Chem. **270**(17): 10091-10096.

Esnault, C., Heidmann, O., Delebecque, F., Dewannieux, M., Ribet, D., Hance, A. J., Heidmann, T. and Schwartz, O. (2005). "APOBEC3G cytidine deaminase inhibits retrotransposition of endogenous retroviruses." Nature **433**(7024): 430-3.

Esnault, C., Maestre, J. and Heidmann, T. (2000). "Human LINE retrotransposons generate processed pseudogenes." Nat Genet **24**(4): 363-7.

Falconer, D. S. (1989). Variance. In *Introduction to quantitative genetics*. Essex, England, Longman Scientific & Technical**:** 136-138.

Fincham, J. R. S. and Sastry, G. R. K. (1974). "Controlling elements in maize." Ann. Rev. Genet. **8**: 15-50.

Finnegan, D. J. (1997). "Transposable elements: How non-LTR retrotransposons do it." Current Biology **7**(4): R245-R248.

Flood, W. D. and Ruvinsky, A. (2001). "Alternative splicing and expressivity of the Axin(Fu) allele in mice." Heredity **87**(Pt 2): 146-52.

Florine, D. L., Ono, T., Cutler, R. G. and Getz, M. J. (1980). "Regulation of endogenous murine leukemia virus-related nuclear and cytoplasmic RNA complexity in C57BL/6J mice of increasing age." Cancer Res **40**(3): 519-23.

Florl, A. R., Lower, R., Schmitz-Drager, B. J. and Schulz, W. A. (1999). "DNA methylation and expression of LINE-1 and HERV-K provirus sequences in urothelial and renal cell carcinomas." Br J Cancer **80**(9): 1312-21.

Frendo, J.-L., Olivier, D., Cheynet, V., Blond, J.-L., Bouton, O., Vidaud, M., Rabreau, M., Evain-Brion, D. and Mallet, F. (2003). "Direct Involvement of HERV-W Env Glycoprotein in Human Trophoblast Cell Fusion and Differentiation." Mol. Cell. Biol. **23**(10): 3566-3574.

Friedrich, G. A., Hildebrand, J. D. and Soriano, P. (1997). "The Secretory Protein Sec8 Is Required for Paraxial Mesoderm Formation in the Mouse." Developmental Biology **192**(2): 364-374.

Fuks, F. (2005). "DNA methylation and histone modifications: teaming up to silence genes." Current Opinion in Genetics & Development **15**(5): 490-495.

Gartner, K. (1990). "A third component causing random variability beside environment and genotype. A reason for the limited success of a 30 year long effort to standardize laboratory animals?" Lab Anim **24**(1): 71-7.

Gartner, K. and Baunack, E. (1981). "Is the similarity of monozygotic twins due to genetic factors alone?" Nature **292**(5824): 646-7.

Gaubatz, J. W., Arcement, B. and Cutler, R. G. (1991). "Gene expression of an endogenous retrovirus-like element during murine development and aging." Mech Ageing Dev **57**(1): 71-85.

Grover, D., Kannan, K., Brahmachari, S. K. and Mukerji, M. (2005). "ALU-ring elements in the primate genomes." Genetica **124**(2-3): 273-89.

Grover, D., Majumder, P. P., Rao, C. B., Brahmachari, S. K. and Mukerji, M. (2003). "Nonrandom Distribution of Alu Elements in Genes of Various Functional Categories: Insight from Analysis of Human Chromosomes 21 and 22." Mol Biol Evol **20**(9): 1420-1424.

Hata, K. and Sakaki, Y. (1997). "Identification of critical CpG sites for repression of L1 transcription by DNA methylation." Gene **189**(2): 227-234.

Hawkins, T. E., Roes, J., Rees, D., Monkhouse, J. and Moss, S. E. (1999). "Immunological Development and Cardiovascular Function Are Normal in Annexin VI Null Mutant Mice." Mol. Cell. Biol. **19**(12): 8028-8032.

Herbst, H., Sauter, M., Kuhler-Obbarius, C., Loning, T. and Mueller-Lantzsch, N. (1998). "Human endogenous retrovirus (HERV)-K transcripts in germ cell and trophoblastic tumours." Apmis **106**(1): 216-20.

Johnson, J. M., Edwards, S., Shoemaker, D. and Schadt, E. E. (2005). "Dark matter in the genome: evidence of widespread transcription detected by microarray tiling experiments." Trends in Genetics **21**(2): 93-102.

Jurka, J. (2000). "Repbase Update: a database and an electronic journal of repetitive elements." Trends in Genetics **16**(9): 418-420.

Kaetzel, M. A. and Dedman, J. R. (2004). "Annexin VI regulation of cardiac function." Biochem Biophys Res Commun **322**(4): 1171-7.

Kafri, T., Ariel, M., Brandeis, M., Shemer, R., Urven, L., McCarrey, J., Cedar, H. and Razin, A. (1992). "Developmental pattern of gene-specific DNA methylation in the mouse embryo and germ line." Genes Dev **6**(5): 705-14.

Kapranov, P., Drenkow, J., Cheng, J., Long, J., Helt, G., Dike, S. and Gingeras, T. R. (2005). "Examples of the complex architecture of the human transcriptome revealed by RACE and high-density tiling arrays." Genome Res. **15**(7): 987-997.

Karlsson, H., Bachmann, S., Schroder, J., McArthur, J., Torrey, E. F. and Yolken, R. H. (2001). "From the Cover: Retroviral RNA identified in the cerebrospinal fluids and brains of individuals with schizophrenia." PNAS **98**(8): 4634-4639.

Kashkush, K., Feldman, M. and Levy, A. A. (2003). "Transcriptional activation of retrotransposons alters the expression of adjacent genes in wheat." Nat Genet **33**(1): 102-106.

Katayama, S., Tomaru, Y., Kasukawa, T., Waki, K., Nakanishi, M., Nakamura, M., Nishida, H., Yap, C. C., Suzuki, M., Kawai, J. *et al.* (2005). "Antisense Transcription in the Mammalian Transcriptome." Science **309**(5740): 1564-1566.

Kidwell, M. G. and Lisch, D. (1997). "Transposable elements as sources of variation in animals and plants." Proc Natl Acad Sci U S A **94**(15): 7704-11.

Kim, D. D. Y., Kim, T. T. Y., Walsh, T., Kobayashi, Y., Matise, T. C., Buyske, S. and Gabriel, A. (2004). "Widespread RNA Editing of Embedded Alu Elements in the Human Transcriptome." Genome Res. **14**(9): 1719-1725.

Kim, S.-H., Kang, Y.-K., Koo, D.-B., Kang, M.-J., Moon, S.-J., Lee, K.-K. and Han, Y.-M. (2004). "Differential DNA methylation reprogramming of various repetitive sequences in mouse preimplantation embryos." Biochemical and Biophysical Research Communications **324**(1): 58-63.

Kjellman, C., Sjogren, H. O., Salford, L. G. and Widegren, B. (1999). "HERV-F (XA34) is a full-length human endogenous retrovirus expressed in placental and fetal tissues." Gene **239**(1): 99-107.

Krainer, A. R. and Maniatis, T. (1985). "Multiple factors including the small nuclear ribonucleoproteins U1 and U2 are necessary for pre-mRNA splicing in vitro." Cell **42**(3): 725-36.

Krieg, A. M., Gourley, M. F. and Perl, A. (1992). "Endogenous retroviruses: potential etiologic agents in autoimmunity." Faseb J **6**(8): 2537-44.

Kuff, E. L. and Lueders, K. K. (1988). "The intracisternal A-particle gene family: structure and functional aspects." Adv Cancer Res. **51**: 183-276.

Kulski, J. K., Gaudieri, S., Bellgard, M., Balmer, L., Giles, K., Inoko, H. and Dawkins, R. L. (1997). "The evolution of MHC diversity by segmental duplication and transposition of retroelements." J Mol Evol **45**(6): 599-609.

Landry, J. R., Medstrand, P. and Mager, D. L. (2001). "Repetitive elements in the 5' untranslated region of a human zinc-finger gene modulate transcription and translation efficiency." Genomics **76**(1-3): 110-6.

Lane, N., Dean, W., Erhardt, S., Hajkova, P., Surani, A., Walter, J. and Reik, W. (2003). "Resistance of IAPs to methylation reprogramming may provide a mechanism for epigenetic inheritance in the mouse." Genesis **35**(2): 88-93.

Lavie, L., Kitova, M., Maldener, E., Meese, E. and Mayer, J. (2005). "CpG Methylation Directly Regulates Transcriptional Activity of the Human Endogenous Retrovirus Family HERV-K(HML-2)." J. Virol. **79**(2): 876-883.

Leib-Mosch, C., Haltmeier, M., Werner, T., Geigl, E.-M., Brack-Werner, R., Francke, U., Erfle, V. and Hehlmann, R. (1993). "Genomic Distribution and Transcription of Solitary HERV-K LTRs." Genomics **18**(2): 261-269.

Leiter, E. H., Fewell, J. W. and Kuff, E. L. (1986). "Glucose induces intracisternal type A retroviral gene transcription and translation in pancreatic beta cells." J Exp Med **163**(1): 87-100.

Lev-Maor, G., Sorek, R., Shomron, N. and Ast, G. (2003). "The birth of an alternatively spliced exon: 3' splice-site selection in Alu exons." Science **300**(5623): 1288-91.

Ling, T.-Y., Chen, C.-L., Huang, Y.-H., Liu, I.-H., Huang, S. S. and Huang, J. S. (2004). "Identification and Characterization of the Acidic pH Binding Sites for Growth Regulatory Ligands of Low Density Lipoprotein Receptor-related Protein-1." J. Biol. Chem. **279**(37): 38736-38748.

Lippman, Z. and Martienssen, R. (2004). "The role of RNA interference in heterochromatic silencing." Nature **431**(7006): 364-370.

Liu, G., Program, N. C. S., Zhao, S., Bailey, J. A., Sahinalp, S. C., Alkan, C., Tuzun, E., Green, E. D. and Eichler, E. E. (2003). "Analysis of Primate Genomic Variation Reveals a Repeat-Driven Expansion of the Human Genome." Genome Res. **13**(3): 358-368.

Mager, D. L., Hunter, D. G., Schertzer, M. and Freeman, J. D. (1999). "Endogenous retroviruses provide the primary polyadenylation signal for two new human genes (HHLA2 and HHLA3)." Genomics **59**(3): 255-63.

Malik, H. S. and Eickbush, T. H. (2001). "Phylogenetic analysis of ribonuclease H domains suggests a late, chimeric origin of LTR retrotransposable elements and retroviruses." Genome Res **11**(7): 1187-97.

Marguerat, S., Wang, W. Y. S., Todd, J. A. and Conrad, B. (2004). "Association of Human Endogenous Retrovirus K-18 Polymorphisms With Type 1 Diabetes." Diabetes **53**(3): 852-854.

Marra, M., Hillier, L., Kucaba, T., Allen, M., Barstead, R., Beck, C., Blistain, A., Bonaldo, M., Bowers, Y., Bowles, L. *et al.* (1999). "An encyclopedia of mouse genes." Nat Genet **21**(2): 191-194.

Martin, N., Boomsma, D. and Machin3, G. (1997). "A twin-pronged attack on complex traits." Nat Genet **17**(4): 387-392.

McHaffie, G. S. and Ralston, S. H. (1995). "Origin of a negative calcium response element in an ALU-repeat: Implications for regulation of gene expression by extracellular calcium." Bone **17**(1): 11-14.

Menendez, L., Benigno, B. B. and McDonald, J. F. (2004). "L1 and HERV-W retrotransposons are hypomethylated in human ovarian carcinomas." Mol Cancer **3**: 12.

Monk, M. (1995). "Epigenetic programming of differential gene expression in development and evolution." Dev Genet **17**(3): 188-97.

Morales, J. F., Snow, E. T. and Murnane, J. P. (2002). "Environmental factors affecting transcription of the human L1 retrotransposon. I. Steroid hormone-like agents." Mutagenesis **17**(3): 193-200.

Moran, J. V., DeBerardinis, R. J. and Kazazian, H. H., Jr. (1999). "Exon shuffling by L1 retrotransposition." Science **283**(5407): 1530-4.

Morgan, H. D., Sutherland, H. G., Martin, D. I. and Whitelaw, E. (1999). "Epigenetic inheritance at the agouti locus in the mouse." Nat Genet **23**(3): 314-8.

Muotri, A. R., Chu, V. T., Marchetto, M. C. N., Deng, W., Moran, J. V. and Gage, F. H. (2005). "Somatic mosaicism in neuronal precursor cells mediated by L1 retrotransposition." **435**(7044): 903-910.

Murdoch, G. H., Sklaviadis, T., Manuelidis, E. E. and Manuelidis, L. (1990). "Potential retroviral RNAs in Creutzfeldt-Jakob disease." J Virol **64**(4): 1477-86.

Naas, T. P., DeBerardinis, R. J., Moran, J. V., Ostertag, E. M., Kingsmore, S. F., Seldin, M. F., Hayashizaki, Y., Martin, S. L. and Kazazian, H. H. (1998). "An actively retrotransposing, novel subfamily of mouse L1 elements." Embo J **17**(2): 590-7.

Nakagawa, K., Brusic, V., McColl, G. and Harrison, L. C. (1997). "Direct evidence for the expression of multiple endogenous retroviruses in the synovial compartment in rheumatoid arthritis." Arthritis Rheum **40**(4): 627-38.

Nigumann, P., Redik, K., Matlik, K. and Speek, M. (2002). "Many human genes are transcribed from the antisense promoter of L1 retrotransposon." Genomics **79**(5): 628-34.

Norton, J. D. and Hogan, B. L. M. (1988). "Temporal and tissue-specific expression of distinct retrovirus-like (VL30) elements during mouse development." Developmental Biology **125**(1): 226-228.

Oei, S.-L., Babich, V. S., Kazakov, V. I., Usmanova, N. M., Kropotov, A. V. and Tomilin, N. V. (2004). "Clusters of regulatory signals for RNA polymerase II transcription associated with Alu family repeats and CpG islands in human promoters." Genomics **83**(5): 873-882.

Okazaki, Y., Furuno, M., Kasukawa, T., Adachi, J., Bono, H., Kondo, S., Nikaido, I., Osato, N., Saito, R., Suzuki, H. *et al.* (2002). "Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs." Nature **420**(6915): 563-573.

Ono, R., Nakamura, K., Inoue, K., Naruse, M., Usami, T., Wakisaka-Saito, N., Hino, T., Suzuki-Migishima, R., Ogonuki, N., Miki, H. *et al.* (2006). "Deletion of Peg10, an imprinted gene acquired from a retrotransposon, causes early embryonic lethality." Nat Genet **38**(1): 101-106.

Ostertag, E. M. and Kazazian, H. H., Jr. (2001). "Biology of mammalian L1 retrotransposons." Annu Rev Genet **35**: 501-38.

Packer, A. I., Manova, K. and Bachvarova, R. F. (1993). "A Discrete LINE-1 Transcript in Mouse Blastocysts." Developmental Biology **157**(1): 281-283.

Pandey, A. and Lewitter, F. (1999). "Nucleotide sequence databases: a gold mine for biologists." Trends in Biochemical Sciences **24**(7): 276-280.

Parkhurst, S. M. and Corces, V. G. (1986). "Interactions among the gypsy transposable element and the yellow and the suppressor of hairy-wing loci in Drosophila melanogaster." Mol Cell Biol **6**(1): 47-53.

Peaston, A. E., Evsikov, A. V., Graber, J. H., de Vries, W. N., Holbrook, A. E., Solter, D. and Knowles, B. B. (2004). "Retrotransposons regulate host genes in mouse oocytes and preimplantation embryos." Dev Cell **7**(4): 597-606.

Pedersen, A. G., Baldi, P., Chauvin, Y. and Brunak, S. (1999). "The biology of eukaryotic promoter prediction--a review." Computers & Chemistry **23**(3-4): 191-207.

Piko, L., Hammons, M. D. and Taylor, K. D. (1984). "Amounts, synthesis, and some properties of intracisternal A particle-related RNA in early mouse embryos." Proc Natl Acad Sci U S A **81**(2): 488-92.

Poznanski, A. A. and Calarco, P. G. (1991). "The expression of intracisternal A particle genes in the preimplantation mouse embryo." Dev Biol **143**(2): 271-81.

Puech, A., Dupressoir, A., Loireau, M. P., Mattei, M. G. and Heidmann, T. (1997). "Characterization of two age-induced intracisternal A-particle-related transcripts in the mouse liver. Transcriptional read-through into an open reading frame with similarities to the yeast ccr4 transcription factor." J Biol Chem **272**(9): 5995-6003.

Rakyan, V. K., Chong, S., Champ, M. E., Cuthbert, P. C., Morgan, H. D., Luu, K. V. and Whitelaw, E. (2003). "Transgenerational inheritance of epigenetic states at the murine Axin(Fu) allele occurs after maternal and paternal transmission." Proc Natl Acad Sci U S A **100**(5): 2538-43.

Risch, N. J. (2000). "Searching for genetic determinants in the new millennium." Nature **405**(6788): 847-56.

Ristevski, S., Purcell, D. F. J., Marshall, J., Campagna, D., Nouri, S., Fenton, S. P., McPhee, D. A. and Kannourakis, G. (1999). "Novel Endogenous Type D Retroviral Particles Expressed at High Levels in a SCID Mouse Thymic Lymphoma." J. Virol. **73**(6): 4662-4669.

Roman-Gomez, J., Jimenez-Velasco, A., Agirre, X., Cervantes, F., Sanchez, J., Garate, L., Barrios, M., Castillejo, J. A., Navarro, G., Colomer, D. *et al.* (2005). "Promoter hypomethylation of the LINE-1 retrotransposable elements activates sense//antisense transcription and marks the progression of chronic myeloid leukemia." Oncogene **24**(48): 7213-7223.

Rote, N. S., Chakrabarti, S. and Stetzer, B. P. (2004). "The role of human endogenous retroviruses in trophoblast differentiation and placental development." Placenta **25**(8-9): 673-683.

Rougier, N., Bourc'his, D., Gomes, D. M., Niveleau, A., Plachot, M., Paldi, A. and Viegas-Pequignot, E. (1998). "Chromosome methylation patterns during mammalian preimplantation development." Genes Dev. **12**(14): 2108-2113.

Rubin, C. M., Kimura, R. H. and Schmid, C. W. (2002). "Selective stimulation of translational expression by Alu RNA." Nucl. Acids Res. **30**(14): 3253-3261.

Russell, P. J. (1998). Genetics of cancer. In *Genetics*. Menlo Park, CA, The Benjamin/Cummings Publishing Company, Inc.**:** 593.

Sans, N., Prybylowski, K., Petralia, R. S., Chang, K., Wang, Y.-X., Racca, C., Vicini, S. and Wenthold, R. J. (2003). "NMDA receptor trafficking through an interaction between PDZ proteins and the exocyst complex." Nat Cell Biol **5**(6): 520-530.

Scadden, A. D. and Smith, C. W. (2001). "RNAi is antagonized by A-->I hyper-editing." EMBO Rep **2**(12): 1107-11.

Schiff, R., Itin, A. and Keshet, E. (1991). "Transcriptional activation of mouse retrotransposons in vivo: specific expression in steroidogenic cells in response to trophic hormones." Genes Dev **5**(4): 521-32.

Schramke, V. and Allshire, R. (2004). "Those interfering little RNAs! Silencing and eliminating chromatin." Current Opinion in Genetics & Development **14**(2): 174-180.

Schultz, J., Doerks, T., Ponting, C. P., Copley, R. R. and Bork, P. (2000). "More than 1,000 putative new human signalling proteins revealed by EST data mining." Nat Genet **25**(2): 201-4.

Sciamanna, I., Landriscina, M., Pittoggi, C., Quirino, M., Mearelli, C., Beraldi, R., Mattei, E., Serafino, A., Cassano, A., Sinibaldi-Vallebona, P. *et al.* (2005). "Inhibition of endogenous reverse transcriptase antagonizes human tumor growth." Oncogene **24**(24): 3923-3931.

Seidl, C., Donner, H., Petershofen, E., Usadel, K. H., Seifried, E., Kaltwasser, J. P. and Badenhoop, K. (1999). "An endogenous retroviral long terminal repeat at the HLA-DQB1 gene locus confers susceptibility to rheumatoid arthritis." Hum Immunol **60**(1): 63-8.

Seitz, H., Youngson, N., Lin, S.-P., Dalbert, S., Paulsen, M., Bachellerie, J.-P., Ferguson-Smith, A. C. and Cavaille, J. (2003). "Imprinted microRNA genes transcribed antisense to a reciprocally imprinted retrotransposon-like gene." Nat Genet **34**(3): 261-262.

Sinibaldi-Vallebona, P., Lavia, P., Garaci, E. and Spadafora, C. (2006). "A role for endogenous reverse transcriptase in tumorigenesis and as a target in differentiating cancer therapy." Genes Chromosomes Cancer **45**(1): 1-10.

Sirokman, G., Humphries, D. E. and Bing, O. H. (1997). "Endogenous retroviral transcripts in myocytes from spontaneously hypertensive rats." Hypertension **30**(1 Pt 1): 88-93.

Smale, S. T. and Kadonaga, J. T. (2003). "The RNA Polymerase II Core Promoter." Annual Review of Biochemistry **72**(1): 449-479.

Smit, A. F. (1996). "The origin of interspresed repeats in the human genome." Curr Opin Genet Dev **6**(6): 743-8.

Soifer, H. S., Zaragoza, A., Peyvan, M., Behlke, M. A. and Rossi, J. J. (2005). "A potential role for RNA interference in controlling the activity of the human LINE-1 retrotransposon." Nucl. Acids Res. **33**(3): 846-856.

Sorek, R., Ast, G. and Graur, D. (2002). "Alu-Containing Exons are Alternatively Spliced." Genome Res. **12**(7): 1060-1067.

Sorek, R. and Safer, H. M. (2003). "A novel algorithm for computational identification of contaminated EST libraries 10.1093/nar/gkg170." Nucl. Acids Res. **31**(3): 1067-1074.

Speek, M. (2001). "Antisense promoter of human L1 retrotransposon drives transcription of adjacent cellular genes." Mol Cell Biol **21**(6): 1973-85.

Stavenhagen, J. B. and Robins, D. M. (1988). "An ancient provirus has imposed androgen regulation on the adjacent mouse sex-limited protein gene." Cell **55**(2): 247-54.

Suter, C., Martin, D. and Ward, R. (2004). "Hypomethylation of L1 retrotransposons in colorectal cancer and adjacent normal tissue." International Journal of Colorectal Disease **19**(2): 95-101.

Svoboda, P., Stein, P., Anger, M., Bernstein, E., Hannon, G. J. and Schultz, R. M. (2004). "RNAi and expression of retrotransposons MuERV-L and IAP in preimplantation mouse embryos." Developmental Biology **269**(1): 276-285.

Tchenio, T., Casella, J.-F. and Heidmann, T. (2000). "Members of the SRY family regulate the human LINE retrotransposons." Nucl. Acids Res. **28**(2): 411-415.

Temin, H. M. (1992). Origin and general nature of retroviruses. In *The Retroviridae*. Levy, J. A. New York, Plenum Press. **1**.

Trasler, J. M. (2005). "Gamete imprinting: setting epigenetic patterns for the next generation." Reproduction, Fertility and Development **18**(2): 63-69.

Triezenberg, S. J. (1992). Primer extension. In *Current Protocols In Molecular Biology*. Struhl, K. United States of America, John Wiley and Sons Inc.**:** 4.8.1-4.8.3.

Tsuboi, T., Ravier, M. A., Xie, H., Ewart, M.-A., Gould, G. W., Baldwin, S. A. and Rutter, G. A. (2005). "Mammalian Exocyst Complex Is Required for the Docking Step of Insulin Vesicle Exocytosis." J. Biol. Chem. **280**(27): 25565-25570.

van de Lagemaat, L. N., Landry, J.-R., Mager, D. L. and Medstrand, P. (2003). "Transposable elements in mammals promote regulatory variation and diversification of genes with specialized functions." Trends in Genetics **19**(10): 530-536.

Van den Broeck, D., Maes, T., Sauer, M., Zethof, J., De Keukeleire, P., D'Hauw, M., Van Montagu, M. and Gerats, T. (1998). "Transposon Display identifies individual transposable elements in high copy number lines." The Plant Journal **13**(1): 121-129.

Vansant, G. and Reynolds, W. (1995). "The Consensus Sequence of a Major Alu Subfamily Contains a Functional Retinoic Acid Response Element." PNAS **92**(18): 8229-8233.

Vasicek, T. J., Zeng, L., Guan, X. J., Zhang, T., Costantini, F. and Tilghman, S. M. (1997). "Two dominant mutations in the mouse fused gene are the result of transposon insertions." Genetics **147**(2): 777-86.

Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., Smith, H. O., Yandell, M., Evans, C. A., Holt, R. A. *et al.* (2001). "The sequence of the human genome." Science **291**: 1304-1351.

Vidal, F., Mougneau, E., Glaichenhaus, N., Vaigot, P., Darmon, M. and Cuzin, F. (1993). "Coordinated Posttranscriptional Control of Gene Expression by Modular Elements Including Alu-Like Repetitive Sequences." PNAS **90**(1): 208-212.

Vinogradova, Leppik, Kalinina, Zhulidov, Grzeschik and Sverdlov (2002). "Selective Differential Display of RNAs containing interspersed repeats: analysis of changes in the transcription of HERV-K LTRs in germ cell tumors." Molecular Genetics and Genomics **266**(5): 796-805.

Vojta, P. J. and Barrett, J. C. (1995). "Genetic analysis of cellular senescence." Biochim Biophys Acta **1242**(1): 29-41.

Walsh, C. P., Chaillet, J. R. and Bestor, T. H. (1998). "Transcription of IAP endogenous retroviruses is constrained by cytosine methylation." Nat Genet **20**(2): 116-117.

Wang-Johanning, F., Frost, A. R., Jian, B., Azerou, R., Lu, D. W., Chen, D. T. and Johanning, G. L. (2003). "Detecting the expression of human endogenous retrovirus E envelope transcripts in human prostate adenocarcinoma." Cancer **98**(1): 187-97.

Waterston, R. H., Lindblad-Toh, K., Birney, E., Rogers, J., Abril, J. F., Agarwal, P., Agarwala, R., Ainscough, R., Alexandersson, M., An, P. *et al.* (2002). "Initial sequencing and comparative analysis of the mouse genome." Nature **420**(6915): 520-562.

Weiner, A. M. (2002). "SINEs and LINEs: the art of biting the hand that feeds you." Current Opinion in Cell Biology **14**(3): 343-350.

Wentzensen, N., Wilz, B., Findeisen, P., Wagner, R., Dippold, W., von Knebel Doeberitz, M. and Gebert, J. (2004). "Identification of differentially expressed genes in colorectal adenoma compared to normal tissue by suppression subtractive hybridization." Int J Oncol **24**(4): 987-94.

Wheelan, S. J., Aizawa, Y., Han, J. S. and Boeke, J. D. (2005). "Gene-breaking: A new paradigm for human retrotransposon-mediated gene evolution." Genome Res. **15**(8): 1073-1078.

Whitelaw, E. and Martin, D. I. (2001). "Retrotransposons as epigenetic mediators of phenotypic variation in mammals." Nat Genet **27**(4): 361-5.

Williams, J. A., Pappu, S. S. and Bell, J. B. (1988). "Molecular analysis of hybrid dysgenesis-induced derivatives of a P-element allele at the vg locus." Mol Cell Biol **8**(4): 1489-1497.

Wolffe, A. P. and Matzke, M. A. (1999). "Epigenetics: Regulation Through Repression." Science **286**(5439): 481-486.

Woodcock, D. M., Lawler, C. B., Linsenmeyer, M. E., Doherty, J. P. and Warren, W. D. (1997). "Asymmetric methylation in the hypermethylated CpG promoter region of the human L1 retrotransposon." J Biol Chem **272**(12): 7810-6.

Yamazaki, T., Takeda, K., Gotoh, K., Takeshima, H., Akira, S. and Kurosaki, T. (2002). "Essential Immunoregulatory Role for BCAP in B Cell Development and Function." J. Exp. Med. **195**(5): 535-545.

Yang, J. N. and Dudley, J. (1992). "Endogenous Mtv-8 or a closely linked sequence stimulates rearrangement of the downstream V kappa 9 gene." J Immunol **149**(4): 1242-51.

Yang, N., Zhang, L., Zhang, Y. and Kazazian, H. H., Jr (2003). "An important role for RUNX3 in human L1 transcription and retrotransposition." Nucl. Acids Res. **31**(16): 4929-4940.

Yang, Z., Boffelli, D., Boonmark, N., Schwartz, K. and Lawn, R. (1998). "Apolipoprotein(a) Gene Enhancer Resides within a LINE Element." J. Biol. Chem. **273**(2): 891-897.

Yelin, R., Dahary, D., Sorek, R., Levanon, E. Y., Goldstein, O., Shoshan, A., Diber, A., Biton, S., Tamir, Y., Khosravi, R. *et al.* (2003). "Widespread occurrence of antisense transcription in the human genome." Nat Biotechnol **21**(4): 379-86.

Yi, J.-M., Kim, H.-M. and Kim, H.-S. (2004). "Expression of the human endogenous retrovirus HERV-W family in various human tissues and cancer cells." J Gen Virol **85**(5): 1203-1210.

Yi, J.-M., Kim, H.-M. and Kim, H.-S. (2006). "Human endogenous retrovirus HERV-H family in human tissues and cancer cells: expression, identification, and phylogeny." Cancer Letters **231**(2): 228-239.

Yi, J.-M. and Kim, H.-S. (2004). "Expression analysis of endogenous retroviral elements belonging to the HERV-F family from human tissues and cancer cells." Cancer Letters **211**(1): 89-96.

Yoder, J. A., Walsh, C. P. and Bestor, T. H. (1997). "Cytosine methylation and the ecology of intragenomic parasites." Trends Genet **13**(8): 335-40.

Yolken, R. H., Karlsson, H., Yee, F., Johnston-Wilson, N. L. and Torrey, E. F. (2000). "Endogenous retroviruses and schizophrenia." Brain Res Brain Res Rev **31**(2-3): 193-9.

Youngson, N. A., Kocialkowski, S., Peel, N. and Ferguson-Smith, A. C. (2005). "A small family of sushi-class retrotransposon-derived genes in mammals and their relation to genomic imprinting." J Mol Evol **61**(4): 481-90.

Yu, X., Zhu, X., Pi, W., Ling, J., Ko, L., Takeda, Y. and Tuan, D. (2005). "The Long Terminal Repeat (LTR) of ERV-9 Human Endogenous Retrovirus Binds to NF-Y in the Assembly of an Active LTR Enhancer Complex NF-Y/MZF1/GATA-2." J. Biol. Chem. **280**(42): 35184-35194.

Zheng, C. L., Fu, X.-D. and Gribscov, M. (2005). "Characteristics and regulatory elements defining constitutive splicing and different modes of alternative splicing in human and mouse." RNA **11**(12): 1777-1787.

APPENDIX

The following Tables contain up to 10 examples of sequences of each type of retroelement, which were obtained from the human and mouse *in silico* dbEST searches. Only a representative sample has been included due to the large number of sequences which were recovered. Dev. Stage is Developmental Stage, Chr. Loc. is Chromosomal location, – indicates the data was not available, #N/A indicates the library details could not be located and ? indicates the chromosomal location could not be determined.

IN SILICO HUMAN RETROELEMENT TRANSCRIPTS

ACCESSION NUMBER	CLONE LIBRARY	TISSUE	CELL TYPE	DEV. STAGE	SEX	DISEASE STATE	CHR. LOC.
ERV-9 Sense Transcripts Starting in LTR							
BE274264	NIH_MGC_20	skin	cell line	-	-	melanotic melanoma	1
BG720667	NIH_MGC_97	testis	-	-	male	normal	10
AU124681	NT2RM4	nervous tissue	NT2 neuronal precursor cells (uninduced)	-	-	teratocarcinoma	11
BE901475	NIH_MGC_21	placenta	cell line	-	-	choriocarcinoma	12
AL557222	Homo sapiens T CELLS (JURKAT CELL LINE)	lymphoreticular tissue	T-cells	-	-	leukemia	13
N40838	Soares_placenta_8to9weeks_2NbHP8 to9W	placenta	-	placenta, 8-9 weeks post conception	-	normal	14
H41984	Soares breast 3NbHBst	mammary gland	-	adult	female	normal	15
H91193	Soares fetal liver spleen 1NFLS	liver and spleen	-	fetus, 20 week post conception	male	normal	17
BE387954	NIH_MGC_44	uterus (endometrium)	adenocarcinoma cell line	-	female	adenocarcinoma	18
BG571021	NIH_MGC_79	placenta	-	-	-	normal	19
ERV-9 Sense Transcripts Ending in LTR							
AI187857	Soares_testis_NHT	testis	-	-	male	normal	1
AW589238	NCI_CGAP_CML1	blood	myeloid cells	-	-	leukemia	10
AA826683	NCI_CGAP_Kid6	kidney	-	-	-	carcinoma	11
AI684924	Soares_NFL_T_GBC_S1	pooled (testis/lung/B-cell)	-	fetus	-	normal	12
AI683611	NCI_CGAP_Ut1	uterus	-	-	female	adenocarcinoma	16
N76435	Soares fetal liver spleen 1NFLS	liver and spleen	-	fetus, 20 week post conception	male	normal	17

AI796860	NCI_CGAP_Lu24	lung	carcinoid	-	-	carcinoma	18
AI808281	Soares_NFL_T_GBC_S1	pooled (testis/lung/B-cell)	-	fetus	-	normal	2
BF431499	NCI_CGAP_Kid11	kidney	-	-	-	normal	20
BF436298	Soares_NSF_F8_9W_OT_PA_P_S1	pooled (parathyroid/ovary/fibroblast/placenta	-	fetus	-	normal	5
ERV-9 Antisense Transcripts Starting in LTR							
AV712881	DCA	uncharacterized tissue	dendritic cells	adult	-	normal	19
R82582	Soares placenta Nb2HP	placenta	-	placenta obtained at birth (full term)	female	normal	20
AA377545	Synovial membrane	synovial membrane	-	adult, 70 year old	male	normal	20
AA286779	NCI_CGAP_GCB1	tonsil	B-cells, germinal centre	-	-	normal	6
H19635	Soares adult brain N2b5HB55Y	brain	pooled	adult, 55 year old	male	ruptured aortic aneurysm	6
ERV-9 Antisense Transcripts Ending in LTR							
AI367188	NCI_CGAP_Ut4	uterus	-	-	female	adenocarcinoma	11
AI825331	NCI_CGAP_GC6	-	germ cells	-	-	tumour (pooled)	17
AA129671	Stratagene lung carcinoma 937218	lung	small cells	cell line	-	carcinoma	2
AV655302	GLC	liver	-	adult	-	normal	5
AA169157	Stratagene fetal retina 937202	eye (retina)	-	fetus	-	normal	6
AA136575	Stratagene lung carcinoma 937218	lung	small cells	cell line	-	carcinoma	X
ERV-9 Unknown Direction							
BE242820	Pediatric acute myelogenous leukemia cell (FAB M1) Baylor-HGSC project=TCAA	bone marrow	myeloid cell	infant, 6 years	male	acute myelogenous leukemia	1
BE146405	HT0209	tongue	-	adult	-	carcinoma	12

BE243895	Pediatric pre-B cell acute lymphoblastic leukemia Baylor-HGSC project=TCBA	lymph node	pre-B cell	infant, 2 years	male	acute lymphoblastic leukemia	14
BE143157	HT0159	#N/A	#N/A	#N/A	#N/A	#N/A	17
AW797835	UM0042	#N/A	#N/A	#N/A	#N/A	#N/A	19
BF175538	Myeloma (MYE) cDNA library	bone marrow	-	-	male	multiple myeloma	2
X98427	Human mRNA (G. La Mantia)	#N/A	#N/A	#N/A	#N/A	#N/A	21
BF883830	ET0211	lung	-	adult	-	carcinoma	3
X98426	Human mRNA (G. La Mantia)	#N/A	#N/A	#N/A	#N/A	#N/A	4
X98428	Human mRNA (G. La Mantia)	#N/A	#N/A	#N/A	#N/A	#N/A	5
HERV-E Sense Transcripts Starting in LTR							
AA299990	Uterus tumor I	uterus	-	adult	female	tumor (neoplasia)	4
AL049113	434 (synonym: htes3)	testis	-	adult	male	normal	6
AL524147	Homo sapiens NEUROBLASTOMA COT 25-NORMALIZED	nervous tissue	-	-	-	neuroblastoma	7
BE540624	NIH_MGC_10	cervix	MGC36	-	female	choriocarcinoma	8
AL529907	Homo sapiens NEUROBLASTOMA COT 50-NORMALIZED	nervous tissue	-	-	-	neuroblastoma	?
T55030	Stratagene fetal spleen (#937205)	spleen	-	fetus	-	normal	?
AI820904	Soares ovary tumor NbHOT	ovary	-	-	female	tumor (neoplasia)	?
BF034544	NIH_MGC_66	ovary	cell line	-	female	adenocarcinoma	?
BG572808	NIH_MGC_79	placenta	-	-	-	normal	?
HERV-E Sense Transcripts Ending in LTR							
AI401637	NCI_CGAP_Pr28	prostate gland	-	adult	male	normal	10
AI091352	Soares_NSF_F8_9W_OT_PA_P_S1	pooled (parathyroid/ovary/fibroblast/placenta	-	fetus	-	normal	11
AI889536	NCI_CGAP_Ut1	uterus	-	-	female	adenocarcinoma	14
AA903655	NCI_CGAP_GC4	-	germ cells	-	-	tumour (pooled)	16
AW874032	NCI_CGAP_Thy3	thyroid	-	-	-	follicular carcinoma	17
AI276404	NCI_CGAP_Ut3	uterus	-	-	female	adenocarcinoma	19
AA568650	NCI_CGAP_Co10	colon	-	-	-	carcinoma	2
AI597706	NCI_CGAP_Pr28	prostate gland	-	adult	male	normal	20
AA758201	Soares_testis_NHT	testis	-	-	male	normal	22

AI243306	Soares_NFL_T_GBC_S1	pooled (testis/lung/B-cell)	-	fetus	-	normal	8
HERV-E Unknown Direction							
AW972524	MAGE resequences, MAGL	colon	-	-	-	tumor metastasis	19
AA492210	NCI_CGAP_Pr6	prostate gland	-	-	male	preneoplasia	20
HERV-F Sense Transcripts Starting in LTR							
R69938	Soares placenta Nb2HP	placenta	-	placenta obtained at birth (full term)	female	normal	7
BG571861	NIH_MGC_79	placenta	-	-	-	normal	7
R32518	Soares placenta Nb2HP	placenta	-	placenta obtained at birth (full term)	female	normal	7
H12467	Soares placenta Nb2HP	placenta	-	placenta obtained at birth (full term)	female	normal	7
R63675	Soares placenta Nb2HP	placenta	-	placenta obtained at birth (full term)	female	normal	7
H79604	Soares fetal liver spleen 1NFLS	liver and spleen	-	fetus, 20 week post conception	male	normal	7
R71045	Soares placenta Nb2HP	placenta	-	placenta obtained at birth (full term)	female	normal	7
R28309	Soares placenta Nb2HP	placenta	-	placenta obtained at birth (full term)	female	normal	7

H78098	Soares fetal liver spleen 1NFLS	liver and spleen	-	fetus, 20 week post conception	male	normal	7
H68658	Soares fetal liver spleen 1NFLS	liver and spleen	-	fetus, 20 week post conception	male	normal	7
HERV-F Sense Transcripts Ending in LTR							
R32705	Soares placenta Nb2HP	placenta	-	placenta obtained at birth (full term)	female	normal	?
R28382	Soares placenta Nb2HP	placenta	-	placenta obtained at birth (full term)	female	normal	?
R26740	Soares placenta Nb2HP	placenta	-	placenta obtained at birth (full term)	female	normal	?
R26685	Soares placenta Nb2HP	placenta	-	placenta obtained at birth (full term)	female	normal	?
R79115	Soares placenta Nb2HP	placenta	-	placenta obtained at birth (full term)	female	normal	?
H13421	Soares placenta Nb2HP	placenta	-	placenta obtained at birth (full term)	female	normal	?
R69804	Soares placenta Nb2HP	placenta	-	placenta obtained at birth (full term)	female	normal	?

R28193	Soares placenta Nb2HP	placenta	-	placenta obtained at birth (full term)	female	normal	?
R24254	Soares placenta Nb2HP	placenta	-	placenta obtained at birth (full term)	female	normal	?
R67876	Soares placenta Nb2HP	placenta	-	placenta obtained at birth (full term)	female	normal	?
HERV-F Unknown Direction							
BG006144	GN0247	placenta	-	adult	-	normal	7
BF990336	GN0160	placenta	-	adult	-	normal	7
AA367885	Placenta I	placenta	-	placenta	-	normal	
HERV-H Sense Transcripts Starting in LTR							
BE866598	NIH_MGC_53	bladder	carcinoma cell line	-	-	carcinoma	1
BE786347	NIH_MGC_68	lung	large cell line	-	-	large cell carcinoma	10
BG771211	NIH_MGC_60	prostate gland	cell line	-	male	adenocarcinoma	11
AA227848	Stratagene NT2 neuronal precursor 937230	brain	Ntera-2 neuroepithelial cells	-	-	normal	12
AA233352	Stratagene NT2 neuronal precursor 937230	brain	Ntera-2 neuroepithelial cells	-	-	normal	13
BF696884	NIH_MGC_56	brain	primitive neuroectoderm	-	-	normal	15
AA299839	Uterus tumor I	uterus	-	adult	female	tumor (neoplasia)	17
BG497859	NIH_MGC_60	prostate gland	cell line	-	male	adenocarcinoma	18
AU100599	Sugano Homo sapiens cDNA library	uncharacterized tissue	-	-	-	-	2
BG498712	NIH_MGC_60	prostate gland	cell line	-	male	adenocarcinoma	20

HERV-H Sense Transcripts Ending in LTR							
BG236150	Soares_NPBMC	blood (white cells)	lymphocyte	-	-	normal	1
AI709011	Barstead colon HPLRB7	colon	-	adult, 25 year old	male	normal	11
AA219557	Stratagene NT2 neuronal precursor 937230	brain	Ntera-2 neuroepithelial cells	-	-	normal	12
AW590142	NCI_CGAP_GC6	-	germ cells	-	-	tumour (pooled)	13
H54619	Soares fetal liver spleen 1NFLS	liver and spleen	-	fetus, 20 week post conception	male	normal	14
AI094704	NCI_CGAP_Brn23	brain	pooled	-	-	glioblastoma	15
AA243224	Stratagene NT2 neuronal precursor 937230	brain	Ntera-2 neuroepithelial cells	-	-	normal	16
AA419145	Soares ovary tumor NbHOT	ovary	-	-	female	tumor (neoplasia)	17
R50526	Soares breast 2NbHBst	mammary gland	-	adult	female	normal	19
AI144451	Soares_fetal_heart_NbHH19W	heart	-	fetus, 19 week post conception	-	normal	2
HERV-H Antisense Transcripts Starting in LTR							
AA418333	Soares_NhHMPu_S1	pooled (melanocyte/fetal heart/and pregnant u	-	fetus	-	normal	1
AA102232	Stratagene colon (#937204)	colon	T84 carcinoma cell line	-	-	carcinoma	?
BE906097	NIH_MGC_70	pancreas	cell line	-	-	epithelioid carcinoma	?
HERV-H Unknown Direction							
BG182246	Athersys RAGE Library	uncharacterized tissue	HT1080 cell line	-	-	-	1
AW967555	MAGE resequences, MAGJ	colon	-	-	-	tumor metastasis	2
H88396	WATM1	adipose tissue, white	-	adult	female	normal	4
BG221200	Athersys RAGE Library	uncharacterized tissue	HT1080 cell line	-	-	-	6
AW970651	MAGE resequences, MAGK	colon	-	-	-	tumor metastasis	11

BF773560	IT0039	#N/A	#N/A	#N/A	#N/A	#N/A	12
BG207394	Athersys RAGE Library	uncharacterized tissue	HT1080 cell line	-	-	-	13
AW975853	MAGE resequences, MAGN	colon	-	-	-	tumor metastasis	22
BG954748	CT0660	colon	-	adult	-	-	?
BG988020	HT1162	#N/A	#N/A	#N/A	#N/A	#N/A	?
HERV-I Sense Transcripts Ending in LTR							
BG287911	NIH_MGC_93	bladder	cell line	-	-	transitional cell papilloma	3
AA017263	Soares retina N2b4HR	eye (retina)	-	adult, 55 year old	male	normal	7
AW269201	NCI_CGAP_Kid11	kidney	-	-	-	normal	13
AI187828	Soares_testis_NHT	testis	-	-	male	normal	21
BG054783	NCI_CGAP_Brn23	brain	pooled	-	-	glioblastoma	22
BF432068	Soares_NSF_F8_9W_OT_PA_P_S1	pooled (parathyroid/ovary/fibroblast/placenta	-	fetus	-	normal	?
AA021273	Soares retina N2b4HR	eye (retina)	-	adult, 55 year old	male	normal	?
BF000987	NCI_CGAP_Br16	mammary gland	-	adult	female	lobullar carcinoma	?
BF590347	Soares_NSF_F8_9W_OT_PA_P_S1	pooled (parathyroid/ovary/fibroblast/placenta	-	fetus	-	normal	X
AV739245	CB	blood (umbilical cord)	-	-	-	normal	Y
HERV-I Unknown Direction							
AW954223	MAGE resequences, MAGC	colon	-	-	-	tumor metastasis	22
BI051447	GN0332	placenta	-	adult	-	normal	?
BG940513	Hembase; Erythroid Progenitor Cells (LCB:ax library)	#N/A	#N/A	#N/A	#N/A	#N/A	7

HERV-P Sense Transcripts Ending in LTR							
AI493740	NCI_CGAP_CLL1	lymphatic system	B-cells	-	-	chronic lymphotic leukemia	?
AI680500	NCI_CGAP_Ut3	uterus	-	-	female	adenocarcinoma	1
AI694241	NCI_CGAP_Co3	colon	-	-	-	tumor (neoplasia)	10
AI140580	Soares_fetal_liver_spleen_1NFLS_S1	liver and spleen	-	fetus, 20 week post conception	male	normal	11
AA825906	NCI_CGAP_GCB1	tonsil	B-cells, germinal centre	-	-	normal	12
AA878672	NCI_CGAP_Kid5	kidney	clear cells	-	-	tumor (2 pooled)	17
AI673317	NCI_CGAP_GC6	-	germ cells	-	-	tumour (pooled)	2
AA431991	Soares_testis_NHT	testis	-	-	male	normal	22
F02044	normalized infant brain cDNA	#N/A	#N/A	#N/A	#N/A	#N/A	3
BE856998	Soares_NSF_F8_9W_OT_PA_P_S1	pooled (parathyroid/ovary/fibroblast/placenta	-	fetus	-	normal	4
HERV-P Antisense Transcripts Starting in LTR							
AA176097	Stratagene neuroepithelium (#937231)	neuroepithelium	Ntera-2/RA neuroepithelial cells	-	-	normal	7
HERV-P Antisense Transcripts Ending in LTR							
AI911374	Soares_NFL_T_GBC_S1	pooled (testis/lung/B-cell)	-	fetus	-	normal	16
AA613237	NCI_CGAP_Phe1	#N/A	#N/A	#N/A	#N/A	#N/A	?
HERV-P Unknown Direction							
AW975806	MAGE resequences, MAGN	colon	-	-	-	tumor metastasis	12
BG184154	Athersys RAGE Library	uncharacterized tissue	HT1080 cell line	-	-	-	12
HERV-R Sense Transcripts Ending in LTR							
AI479940	NCI_CGAP_Brn25	brain	-	-	-	anaplastic oligodendroglioma	7
AW170423	Soares_NHCeC_cervical_tumor	cervix	-	-	female	carcinoma	7

AA469075	NCI_CGAP_Co3	colon	-	-	-	tumor (neoplasia)	7
AI868243	Soares_total_fetus_Nb2HF8_9w	fetus (total)	-	fetus, 8-9 weeks	-	normal	7
AA634137	Stratagene hNT neuron (#937233)	nervous tissue	hNT neurons (differentiated, post mitotic)	-	-	normal	7
AW025051	NCI_CGAP_Kid3	kidney	-	-	-	normal	7
AI078475	Soares_fetal_liver_spleen_1NFLS_S1	liver and spleen	-	fetus, 20 week post conception	male	normal	7
AW138514	NCI_CGAP_Sub3	pooled	-	-	-	-	7
BF110863	NCI_CGAP_Lu24	lung	carcinoid	-	-	carcinoma	7
N32401	Soares_placenta_8to9weeks_2NbHP8 to9W	placenta	-	placenta, 8-9 weeks post conception	-	normal	7
HERV-R Unknown Direction							
AA027762	HPLA CCLee	#N/A	#N/A	#N/A	#N/A	#N/A	7
HERV-W Sense Transcripts Starting in LTR							
AV708482	ADC	adrenal gland	-	adult	-	normal	15
BF308839	NIH_MGC_17	muscle	cell line	-	-	rhabdomyosarcoma	7
BI087653	NIH_MGC_10	cervix	MGC36	-	female	choriocarcinoma	7
R77331	Soares placenta Nb2HP	placenta	-	placenta obtained at birth (full term)	female	normal	7
BE730884	NIH_MGC_21	placenta	cell line	-	-	choriocarcinoma	7
R27689	Soares placenta Nb2HP	placenta	-	placenta obtained at birth (full term)	female	normal	7
T87403	Soares fetal liver spleen 1NFLS	liver and spleen	-	fetus, 20 week post conception	male	normal	7

BG573364	NIH_MGC_79	placenta	-	-	-	normal	7
BI087886	NIH_MGC_10	cervix	MGC36	-	female	choriocarcinoma	7
HERV-W Sense Transcripts Ending in LTR							
AA868777	Soares_testis_NHT	testis	-	-	male	normal	1
AA250958	NCI_CGAP_GCB1	tonsil	B-cells, germinal centre	-	-	normal	10
R27412	Soares placenta Nb2HP	placenta	-	placenta obtained at birth (full term)	female	normal	14
AI003607	Soares_pineal_gland_N3HPG	pineal gland	-	-	-	normal	19
AA774109	Stratagene hNT neuron (#937233)	nervous tissue	hNT neurons (differentiated, post mitotic)	-	-	normal	4
AI611840	NCI_CGAP_Brn52	brain	-	-	-	astrocytoma/meningioma/ oligodendroglioma/medull oblastoma	5
R76086	Soares placenta Nb2HP	placenta	-	placenta obtained at birth (full term)	female	normal	7
AI828489	NCI_CGAP_Pan1	pancreas	-	-	-	adenocarcinoma	9
AI288235	Soares_NhHMPu_S1	pooled (melanocyte/fetal heart/and pregnant u	-	fetus	-	normal	?
AI379210	NCI_CGAP_CLL1	lymphatic system	B-cells	-	-	chronic lymphotic leukemia	X
HERV-W Unknown Direction							
BI022788	MT0236	#N/A	#N/A	#N/A	#N/A	#N/A	10
AW971553	MAGE resequences, MAGL	colon	-	-	-	tumor metastasis	10
AA837267	NCI_CGAP_GCB1	tonsil	B-cells, germinal centre	-	-	normal	10
AA729556	NCI_CGAP_Alv1	lung	-	-	-	alveolar rhabdomyosarcoma	7

BE247592	Pediatric pre-B cell acute lymphoblastic leukemia Baylor-HGSC project=TCBA	lymph node	pre-B cell	infant, 2 years	male	acute lymphoblastic leukemia	Y
RRHERV-I Sense Transcripts Ending in LTR							
AA938285	NCI_CGAP_Kid5	kidney	clear cells	-	-	tumor (2 pooled)	1
AA897271	Soares_NFL_T_GBC_S1	pooled (testis/lung/B-cell)	-	fetus	-	normal	11
AW511910	NCI_CGAP_Kid8	kidney	renal cells	-	-	renal cell tumor	14
AI684569	Soares_NFL_T_GBC_S1	pooled (testis/lung/B-cell)	-	fetus	-	normal	16
AA017402	Soares retina N2b4HR	eye (retina)	-	adult, 55 year old	male	normal	19
N34500	Soares_multiple_sclerosis_2NbHMSP	brain	multiple sclerosis lesions	adult, 46 year old	male	multiple sclerosis	2
W15274	Soares_parathyroid_tumor_NbHPA	parathyroid gland	-	adult	-	adenoma	3
N59469	Soares_multiple_sclerosis_2NbHMSP	brain	multiple sclerosis lesions	adult, 46 year old	male	multiple sclerosis	5
R02550	Soares fetal liver spleen 1NFLS	liver and spleen	-	fetus, 20 week post conception	male	normal	8
AI825473	NCI_CGAP_GC6	-	germ cells	-	-	tumour (pooled)	9
RRHERV-I Antisense Transcripts Starting in LTR							
BE254053	NIH_MGC_16	eye	cell line	-	-	retinoblastoma	18
RRHERV-I Unknown Transcripts							
AA350196	Infant brain	brain	-	infant	female	normal	1
BI027318	MT0291	#N/A	#N/A	#N/A	#N/A	#N/A	16
F15197	STRATAGENE Human skeletal muscle cDNA library, cat. #936215.	skeletal muscle (leg)	-	adult, 19 year old	female	normal	3
AA657610	NCI_CGAP_Alv1	lung	-	-	-	alveolar rhabdomyosarcoma	3
S71 Sense Transcripts Ending in LTR							
AI869639	NCI_CGAP_Brn25	brain	-	-	-	anaplastic oligodendroglioma	?

AA001134	Soares retina N2b4HR	eye (retina)	-	adult, 55 year old	male	normal	1
AI032523	Soares_fetal_liver_spleen_1NFLS_S1	liver and spleen	-	fetus, 20 week post conception	male	normal	10
BF435997	Soares_NSF_F8_9W_OT_PA_P_S1	pooled (parathyroid/ovary/fibroblast/placenta	-	fetus	-	normal	11
AA401541	Soares_testis_NHT	testis	-	-	male	normal	12
AA813699	Soares_testis_NHT	testis	-	-	male	normal	14
AA609495	Soares_testis_NHT	testis	-	-	male	normal	19
AI371706	Soares_total_fetus_Nb2HF8_9w	fetus (total)	-	fetus, 8-9 weeks	-	normal	21
AI202627	NCI_CGAP_Pr28	prostate gland	-	adult	male	normal	3
AI075265	NCI_CGAP_CLL1	lymphatic system	B-cells	-	-	chronic lymphotic leukemia	5
S71 Antisense Transcripts Starting in LTR							
AA376387	HSC172 cells II	lung	fibroblasts	fetus	-	normal	6
S71 Unknown Transcripts							
AA876910	NCI_CGAP_Pr12	prostate gland	-	-	male	metastatic prostate lesion	19
BE065130	BT0314	#N/A	#N/A	#N/A	#N/A	#N/A	6
BE065134	BT0314	#N/A	#N/A	#N/A	#N/A	#N/A	6
AW954418	MAGE resequences, MAGC	colon	-	-	-	tumor metastasis	6
BE693401	BT0314	#N/A	#N/A	#N/A	#N/A	#N/A	6

LINE Sense Transcripts Starting in Promoter							
BF793477	NIH_MGC_84	adrenal gland (cortex)	adrenal cortex carcinoma, cell line	-	-	adrenal cortex carcinoma	1
BG431078	NIH_MGC_75	kidney	-	-	-	normal	2
BF673583	NIH_MGC_83	prostate gland	-	-	male	normal	3
BG541337	NIH_MGC_77	lung	-	-	-	normal	4
BG546115	NIH_MGC_77	lung	-	-	-	normal	5
H63688	Soares fetal liver spleen 1NFLS	liver and spleen	-	fetus, 20 week post conception	male	normal	6
BE728115	NIH_MGC_20	skin	cell line	-	-	melanotic melanoma	7
T57455	Stratagene fetal spleen (#937205)	spleen	-	fetus	-	normal	8
AL133895	761 (synonym: hamy2)	brain (amygdala)	-	adult	-	normal	9
BF217767	NIH_MGC_57	brain	cell line	-	-	glioblastoma	10
LINE Sense Transcripts Ending in Promoter							
AA524729	NCI_CGAP_Co3	colon	-	-	-	tumor (neoplasia)	1
T57668	Stratagene lung (#937210)	lung	-	adult, 72 year old	male	normal	2
AA833581	Soares_testis_NHT	testis	-	-	male	normal	3
AA553442	NCI_CGAP_Sch1	nervous tissue	-	-	-	Schwannoma tumour	4
BF592342	NCI_CGAP_Br16	mammary gland	-	adult	female	lobullar carcinoma	5
AA662993	Stratagene fetal retina 937202	eye (retina)	-	fetus	-	normal	6
AV652938	GLC	liver	-	adult	-	normal	7
AI926740	NCI_CGAP_Gas4	stomach	signet ring cells	-	-	adenocarcinoma	8
T63786	Stratagene lung (#937210)	lung	-	adult, 72 year old	male	normal	9
N62150	Soares_multiple_sclerosis_2NbHMSP	brain	multiple sclerosis lesions	adult, 46 year old	male	multiple sclerosis	10
LINE Antisense Transcripts Starting in Promoter							
AI347154	NCI_CGAP_Co16	colon	-	-	-	carcinoma	1
AA121877	Stratagene fetal retina 937202	eye (retina)	-	fetus	-	normal	2
AV652798	GLC	liver	-	adult	-	normal	3

AI380547	NCI_CGAP_CLL1	lymphatic system	B-cells	-	-	chronic lymphotic leukemia	4
AI014333	Johnston frontal cortex	brain (frontal cortex)	pooled	adult	male	normal	8
AA188644	Stratagene HeLa cell s3 937216	cervix	HeLa S3 cells	-	female	carcinoma	10
AI349954	NCI_CGAP_Lu26	lung	-	adult	-	invasive adenocarcinoma	11
AI000799	NCI_CGAP_Kid3	kidney	-	-	-	normal	16
AA180278	Stratagene fetal retina 937202	eye (retina)	-	fetus	-	normal	18
AA774178	Stratagene hNT neuron (#937233)	nervous tissue	hNT neurons (differentiated, post mitotic)	-	-	normal	21
LINE Antisense Transcripts Ending in Promoter							
BF218444	NIH_MGC_57	brain	cell line	-	-	glioblastoma	1
BI825742	NIH_MGC_119	brain (medulla)	-	-	-	normal	2
BE386373	NIH_MGC_20	skin	cell line	-	-	melanotic melanoma	3
BG420928	NIH_MGC_14	kidney	renal cell	-	-	adenocarcinoma	4
BE568818	NIH_MGC_53	bladder	carcinoma cell line	-	-	carcinoma	5
BG399800	NIH_MGC_75	kidney	-	-	-	normal	6
BG391018	NIH_MGC_92	testis	cell line	embryo	-	embryonal carcinoma	7
BI438437	HR85 islet	pancreas	pancreatic islet	-	-	normal	8
AA137026	Stratagene fetal retina 937202	eye (retina)	-	fetus	-	normal	9
AV731015	HTF	brain (hypothalamus)	-	adult	-	normal	10
LINE Unknown Transcripts							
BG988585	HT0023	tongue	-	adult	-	carcinoma	1
AW899342	NN0087	#N/A	#N/A	#N/A	#N/A	#N/A	2
AA225638	NCI_CGAP_Pr1	#N/A	#N/A	#N/A	#N/A	#N/A	3
BI050408	GN0294	#N/A	#N/A	#N/A	#N/A	#N/A	4
BF874267	ET0145	lung	-	adult	-	carcinoma	5
BF815609	CI0128	colon	cell line	adult	-	adenocarcinoma	6
AA826146	NCI_CGAP_Ov2	ovary	-	-	female	adenocarcinoma	7
AA804919	NCI_CGAP_Pr16	prostate gland	-	-	male	carcinoma	8
BF733347	AN0039	amniotic fluid	fibroblast cell line	adult	-	normal	9
BG009679	GN0191	placenta	-	adult	-	normal	10

Alu Consensus Sense Transcripts Starting in Promoter							
BF244530	NIH_MGC_57	brain	cell line	-	-	glioblastoma	
BG547647	NIH_MGC_77	lung	-	-	-	normal	
BF677892	NIH_MGC_83	prostate gland	-	-	male	normal	
BG570886	NIH_MGC_79	placenta	-	-	-	normal	
AV761107	MDS	bone marrow	-	-	-	normal	
BE883501	NIH_MGC_71	uterus	cell line	-	female	leiomyosarcoma	
BG564103	NIH_MGC_76	liver	-	-	-	normal	
BG529995	NIH_MGC_61	testis	cell line	embryo	male	embryonal carcinoma	
BG527182	NIH_MGC_59	salivary gland	cell line	-	-	mucoepidermoid carcinoma	
AW023111	Morton Fetal Cochlea	ear (cochlea)	-	fetus, 16-22 weeks	-	normal	
Alu Consensus Sense Transcripts Ending in Promoter							
N64587	Morton Fetal Cochlea	ear (cochlea)	-	fetus, 16-22 weeks	-	normal	
BG057207	Lupski_dorsal_root_ganglion	nervous tissue	dorsal root ganglia	adult, 36 year old	male	normal	
BF939954	NCI_CGAP_Brn23	brain	pooled	-	-	glioblastoma	
AW272294	NCI_CGAP_Co14	colon	-	-	-	moderately-differentiated adenocarcinoma	
AI870453	NCI_CGAP_Ut1	uterus	-	-	female	adenocarcinoma	
AW516510	NCI_CGAP_Ov39	ovary	-	-	female	papillary serous carcinoma	
AW438643	NCI_CGAP_Ut4	uterus	-	-	female	adenocarcinoma	
AI344844	NCI_CGAP_Lu26	lung	-	adult	-	invasive adenocarcinoma	
AI610651	NCI_CGAP_Gas4	stomach	signet ring cells	-	-	adenocarcinoma	
AW192331	NCI_CGAP_Pan1	pancreas	-	-	-	adenocarcinoma	
Alu Consensus Antisense Transcripts Starting in Promoter							
BE047069	NCI_CGAP_HN13	tongue	squamous cell	-	-	carcinoma	
AI340453	NCI_CGAP_Brn20	brain	-	adult	-	oligodendroglioma	
AI345654	NCI_CGAP_Lu26	lung	-	adult	-	invasive adenocarcinoma	
AI246080	NCI_CGAP_Ov32	ovary	-	-	female	papillary serous carcinoma	

AW302724	NCI_CGAP_Ov26	ovary	-	adult	female	papillary serous carcinoma	
AA179814	Stratagene fetal retina 937202	eye (retina)	-	fetus	-	normal	
BE677244	Lupski_dorsal_root_ganglion	nervous tissue	dorsal root ganglia	adult, 36 year old	male	normal	
AW302017	NCI_CGAP_Ov26	ovary	-	adult	female	papillary serous carcinoma	
AI612070	NCI_CGAP_HSC4	#N/A	#N/A	#N/A	#N/A	#N/A	
AI611533	NCI_CGAP_HSC4	#N/A	#N/A	#N/A	#N/A	#N/A	
Alu Consensus Antisense Transcripts Ending in Promoter							
BG059314	NCI_CGAP_HN19	nasopharynx	epithelium	-	-	normal	
BF678323	NIH_MGC_83	prostate gland	-	-	male	normal	
AV719506	GLC	liver	-	adult	-	normal	
BG547829	NIH_MGC_77	lung	-	-	-	normal	
W03818	Soares fetal liver spleen 1NFLS	liver and spleen	-	fetus, 20 week post conception	male	normal	
AA350859	Infant brain	brain	-	infant	female	normal	
AA722297	Stratagene hNT neuron (#937233)	nervous tissue	hNT neurons (differentiated, post mitotic)	-	-	normal	
AI475569	NCI_CGAP_CLL1	lymphatic system	B-cells	-	-	chronic lymphotic leukemia	
BG527382	NIH_MGC_59	salivary gland	cell line	-	-	mucoepidermoid carcinoma	
BE294700	NIH_MGC_17	muscle	cell line	-	-	rhabdomyosarcoma	
Alu Consensus Unknown Transcripts							
AW949694	MAGE resequences, MAGA	colon	-	-	-	tumor metastasis	
AA657835	NCI_CGAP_Pr2	prostate gland	preneoplastic	adult, 45 year old	male	proneoplasia	
F13749	Atrium cDNA library Human heart	heart	atrium	-	-	normal	
AA493787	NCI_CGAP_Thy1	thyroid	-	-	-	carcinoma	
AA491814	NCI_CGAP_Lip2	adipose tissue	-	-	-	liposarcoma	
AW813890	ST0197	stomach	-	adult	-	carcinoma	

AI751162	Normal Human Trabecular Bone Cells	bone (hip)	trabecular bone cells	-	female	normal	
AI064864	Human fetal liver cDNA library	liver	-	fetus	-	normal	
AA525824	NCI_CGAP_Ov2	ovary	-	-	female	adenocarcinoma	
AA746681	NCI_CGAP_Alv1	lung	-	-	-	alveolar rhabdomyosarcoma	
AluJo Sense Transcripts Starting in Promoter							
AL119737	761 (synonym: hamy2)	brain (amygdala)	-	adult	-	normal	
BI087347	NIH_MGC_10	cervix	MGC36	-	female	choriocarcinoma	
BG613924	NIH_MGC_61	testis	cell line	embryo	male	embryonal carcinoma	
AU120494	HEMBB1	embryo (mainly body)	-	embryo, 10 weeks	-	normal	
T60666	Stratagene lung (#937210)	lung	-	adult, 72 year old	male	normal	
BF695337	NIH_MGC_81	skeletal muscle	-	-	-	normal	
H67234	Weizmann Olfactory Epithelium	nose	olfactory epithelium	adult, 35 year old	female	normal	
BF984807	NIH_MGC_88	small intestine	cell line	-	-	duodenal adenocarcinoma	
AV762454	MDS	bone marrow	-	-	-	normal	
AU130007	NT2RP3	nervous tissue	NT2 neuronal precursor cells	-	-	teratocarcinoma	
AluJo Sense Transcripts Ending in Promoter							
AA985143	Stratagene schizo brain S11	brain (frontal cortex)	-	adult, 34 year old	male	schizophrenia	
BF591764	NCI_CGAP_Br16	mammary gland	-	adult	female	lobullar carcinoma	
AA599080	Stratagene lung carcinoma 937218	lung	small cells	cell line	-	carcinoma	
AI469586	NCI_CGAP_Co14	colon	-	-	-	moderately-differentiated adenocarcinoma	
AI271762	NCI_CGAP_Kid3	kidney	-	-	-	normal	
AW591746	NCI_CGAP_Ut1	uterus	-	-	female	adenocarcinoma	
AW238495	NCI_CGAP_HN10	head and neck	retromolar trigone	-	-	carcinoma	
AI279417	NCI_CGAP_Ov36	ovary	-	adult	female	borderline ovarian carcinoma	

AW080134	NCI_CGAP_Eso2	esophagus	squamous cells	-	-	squamous cell carcinoma	
AA706628	Soares_parathyroid_tumor_NbHPA	parathyroid gland	-	adult	-	adenoma	
AluJo Antisense Transcripts Starting in Promoter							
AW238214	NCI_CGAP_HN10	head and neck	retromolar trigone	-	-	carcinoma	
AI298462	NCI_CGAP_Lu5	lung	carcinoid	-	-	carcinoma	
N66948	Soares fetal liver spleen 1NFLS	liver and spleen	-	fetus, 20 week post conception	male	normal	
BF942009	NCI_CGAP_Thy10	thyroid	-	-	-	medullary carcinoma	
AW081610	NCI_CGAP_Co18	colon	-	-	-	adenocarcinoma	
AI620992	NCI_CGAP_Gas4	stomach	signet ring cells	-	-	adenocarcinoma	
N67313	Morton Fetal Cochlea	ear (cochlea)	-	fetus, 16-22 weeks	-	normal	
BE139139	NCI_CGAP_Ov26	ovary	-	adult	female	papillary serous carcinoma	
BM023196	Melton Normalized Human Islet 4 N4-HIS 1	Pancreas	Islets of Langerhans	adult	-	normal	
AW064294	KRIBB Human CD4 intrathymic T-cell cDNA library	thymus	CD3+4+8- single positive stage	-	-	normal	
AluJo Antisense Transcripts Ending in Promoter							
AL046225	434 (synonym: htes3)	testis	-	adult	male	normal	
H53284	Soares fetal liver spleen 1NFLS	liver and spleen	-	fetus, 20 week post conception	male	normal	
AV719789	GLC	liver	-	adult	-	normal	
AW672927	NIH_MGC_10	cervix	MGC36	-	female	choriocarcinoma	
BM044552	NIH_MGC_40	prostate gland	cell line	-	male	carcinoma	
C18083	Human placenta cDNA (TFujiwara)	placenta	-	placenta	-	normal	
AA371410	Prostate gland I	prostate gland	-	adult, 21 year old	male	normal	
BG566328	NIH_MGC_76	liver	-	-	-	normal	
AV720667	GLC	liver	-	adult	-	normal	
AL134167	547 (synonym: hfbr1)	brain	-	fetus	-	normal	

AluJo Unknown Transcripts							
BG958534	CT0802	colon	-	adult	-	carcinoma	
AA502451	NCI_CGAP_Lip2	adipose tissue	-	-	-	liposarcoma	
BF988285	GN0163	placenta	-	adult	-	normal	
AA635310	NCI_CGAP_Pr3	prostate gland	malignant cancer cells	adult, 45 year old	male	adenocarcinoma	
AW606809	HT0422	thyroid	-	adult	-	carcinoma	
AA838175	NCI_CGAP_Ov2	ovary	-	-	female	adenocarcinoma	
AA729755	NCI_CGAP_Alv1	lung	-	-	-	alveolar rhabdomyosarcoma	
AF075373	Human fetal liver cDNA library	liver	-	fetus	-	normal	
BF947939	NN0214	brain	-	adult	-	normal	
BE247305	Pediatric pre-B cell acute lymphoblastic leukemia Baylor-HGSC project=TCBA	lymph node	pre-B cell	infant, 2 years	male	acute lymphoblastic leukemia	
AluSc Sense Transcripts Starting in Promoter							
BG433414	NIH_MGC_75	kidney	-	-	-	normal	
AV764104	MDS	bone marrow	-	-	-	normal	
BF680805	NIH_MGC_83	prostate gland	-	-	male	normal	
BG391274	NIH_MGC_92	testis	cell line	embryo	-	embryonal carcinoma	
BG539134	NIH_MGC_77	lung	-	-	-	normal	
AL603463	686 (synonym: hlcc3)	skeletal muscle	-	adult	-	normal	
BF979071	NIH_MGC_62	skin	cell line	-	-	melanotic melanoma, high MDR	
AV734666	cdA	adrenal gland	-	adult	-	pheochromocytoma	
BF697673	NIH_MGC_56	brain	primitive neuroectoderm	-	-	normal	
AV735370	CB	blood (umbilical cord)	-	-	-	normal	
AluSc Sense Transcripts Ending in Promoter							
AI859834	NCI_CGAP_Ut4	uterus	-	-	female	adenocarcinoma	
BE300645	NIH_MGC_8	lymph node	-	-	-	Burkitt lymphoma	
AI345157	NCI_CGAP_Lu26	lung	-	adult	-	invasive adenocarcinoma	
BI711845	Human insulinoma	pancreas	-	-	-	insulinoma	
AI016704	NCI_CGAP_Kid3	kidney	-	-	-	normal	

BF057326	NCI_CGAP_Ov18	ovary	-	-	female	fibrotheoma	
AI445815	NCI_CGAP_Gas4	stomach	signet ring cells	-	-	adenocarcinoma	
AL041706	434 (synonym: htes3)	testis	-	adult	male	normal	
AI281697	NCI_CGAP_Eso2	esophagus	squamous cells	-	-	squamous cell carcinoma	
BG236628	NCI_CGAP_HN20	head and neck	-	-	-	normal	
AluSc Antisense Transcripts Starting in Promoter							
AI305766	NCI_CGAP_Ov33	#N/A	#N/A	#N/A	#N/A	#N/A	
AW302013	NCI_CGAP_Ov26	ovary	-	adult	female	papillary serous carcinoma	
AW467340	NCI_CGAP_CML1	blood	myeloid cells	-	-	leukemia	
AI252506	NCI_CGAP_Ov31	ovary	-	-	female	papillary serous carcinoma	
AW265009	NCI_CGAP_Co22	#N/A	#N/A	#N/A	#N/A	#N/A	
BE676900	NCI_CGAP_Thy11	thyroid	-	-	-	follicular carcinoma	
AluSc Antisense Transcripts Ending in Promoter							
AV718479	GLC	liver	-	adult	-	normal	
AW630298	NCI_CGAP_GU1	genitourinary tract	transitional cell	-	-	carcinoma	
BM010698	NIH_MGC_41	skin	cell line	-	-	amelanotic melanoma	
AA384039	Thyroid	thyroid	-	adult	-	normal	
AV732042	HTF	brain (hypothalamus)	-	adult	-	normal	
BF982266	NIH_MGC_88	small intestine	cell line	-	-	duodenal adenocarcinoma	
BG570559	NIH_MGC_77	lung	-	-	-	normal	
AA437161	Soares_testis_NHT	testis	-	-	male	normal	
BF725347	Human Iris cDNA (Un-normalized, unamplified): BX	eye (iris)	-	adult	-	normal	
AU120769	HEMBB1	embryo (mainly body)	-	embryo, 10 weeks	-	normal	

AluSc Unknown Transcripts							
AW977540	MAGE resequences, MAGO	colon	-	-	-	tumor metastasis	
AW819626	ST0293	stomach	-	adult	-	carcinoma	
F17700	HM1	skeletal muscle (pectoral)	-	-	female	after mastectomy	
AA559290	NCI_CGAP_Pr4	prostate gland	-	adult	male	preneoplasia	
AA528480	NCI_CGAP_Kid1	kidney	renal cells	-	-	renal cell carcinoma	
BF919090	NT0135	brain	-	adult	-	carcinoma	
AA526193	NCI_CGAP_Ov2	ovary	-	-	female	adenocarcinoma	
AF074677	Human fetal liver cDNA library	liver	-	fetus	-	normal	
AA482896	NCI_CGAP_Alv1	lung	-	-	-	alveolar rhabdomyosarcoma	
AA610491	NCI_CGAP_Thy1	thyroid	-	-	-	carcinoma	
AluSp Sense Transcripts Starting in Promoter							
BG540416	NIH_MGC_77	lung	-	-	-	normal	
BG180437	NIH_MGC_91	prostate gland	cell line	-	male	adenocarcinoma	
AV763538	MDS	bone marrow	-	-	-	normal	
AA362349	Macrophage II	lymphoreticular tissue	macrophage	adult	-	normal	
BG621328	NIH_MGC_79	placenta	-	-	-	normal	
BF248297	NIH_MGC_62	skin	cell line	-	-	melanotic melanoma, high MDR	
BG431299	NIH_MGC_75	kidney	-	-	-	normal	
BG654402	HR85 islet	pancreas	pancreatic islet	-	-	normal	
AV737931	CB	blood (umbilical cord)	-	-	-	normal	
BE253048	NIH_MGC_16	eye	cell line	-	-	retinoblastoma	
AluSp Sense Transcripts Ending in Promoter							
AI537538	NCI_CGAP_Gas4	stomach	signet ring cells	-	-	adenocarcinoma	
AI587583	NCI_CGAP_Pan1	pancreas	-	-	-	adenocarcinoma	
AI587565	NCI_CGAP_Pan1	pancreas	-	-	-	adenocarcinoma	
AI580707	NCI_CGAP_Gas4	stomach	signet ring cells	-	-	adenocarcinoma	
BG059523	NCI_CGAP_Brn65	brain	-	-	-	glioblastoma without EGFR amplification	
AA593471	NCI_CGAP_Gas1	#N/A	#N/A	#N/A	#N/A	#N/A	
AA587991	NCI_CGAP_AA1	adrenal gland (cortex)	-	-	-	adrenal adenoma	

R88067	Soares adult brain N2b4HB55Y	brain	-	adult, 55 year old	-	ruptured aortic aneurysm	
AW270258	NCI_CGAP_HN11	tongue	squamous epithelium	-	-	normal	
AluSp Antisense Transcripts Starting in Promoter							
BE676915	NCI_CGAP_Thy11	thyroid	-	-	-	follicular carcinoma	
BE301111	NIH_MGC_14	kidney	renal cell	-	-	adenocarcinoma	
AluSp Antisense Transcripts Ending in Promoter							
BE795275	NIH_MGC_7	lung	MGC3	-	-	carcinoma	
BG393529	NIH_MGC_92	testis	cell line	embryo	-	embryonal carcinoma	
AL596543	451 (synonym: hlcc1) spinal cord	spinal cord	-	adult	-	normal	
AA326441	Cerebellum II	brain (cerebellum)	-	adult	-	normal	
AluSp Unknown Transcripts							
AW963565	MAGE resequences, MAGH	colon	-	-	-	tumor metastasis	
AA862243	NCI_CGAP_Ov2	ovary	-	-	female	adenocarcinoma	
C06046	Human pancreatic islet	pancreas	pancreatic islets	adult	-	normal	
BF854401	EN0092	lung	-	adult	-	normal	
AA640617	NCI_CGAP_Pr2	prostate gland	preneoplastic	adult, 45 year old	male	proneoplasia	
AA610509	NCI_CGAP_Thy1	thyroid	-	-	-	carcinoma	
AA557440	NCI_CGAP_Pr7	prostate gland	-	-	male	low-grade prostatic neoplasia	
BF815810	CI0191	colon	-	adult	-	adenocarcinoma	
AA548029	NCI_CGAP_Pr7	prostate gland	-	-	male	low-grade prostatic neoplasia	
BE177236	HT0592	thyroid	-	adult	-	carcinoma	
AluSx Sense Transcripts Starting in Promoter							
BF978936	NIH_MGC_62	skin	cell line	-	-	melanotic melanoma, high MDR	
BG535971	NIH_MGC_77	lung	-	-	-	normal	
BF216280	NIH_MGC_57	brain	cell line	-	-	glioblastoma	
AV744179	CB	blood (umbilical cord)	-	-	-	normal	

AV760941	MDS	bone marrow	-	-	-	normal	
BF676548	NIH_MGC_83	prostate gland	-	-	male	normal	
AV754716	TP	pituitary gland	-	adult	-	pituitary tumor	
BG430556	NIH_MGC_75	kidney	-	-	-	normal	
AV706891	ADB	adrenal gland	-	adult	-	normal	
BF030666	NIH_MGC_58	kidney	cell line	-	-	hypernephroma	
AV712092	DCA	uncharacterized tissue	dendritic cells	adult	-	normal	
AluSx Sense Transcripts Ending in Promoter							
BF939548	NCI_CGAP_Brn23	brain	pooled	-	-	glioblastoma	
AI917132	NCI_CGAP_Kid8	kidney	renal cells	-	-	renal cell tumor	
AluSx Antisense Transcripts Starting in Promoter							
AW265688	NCI_CGAP_HN11	tongue	squamous epithelium	-	-	normal	
AA714110	NCI_CGAP_SS1	connective tissue	-	-	-	synovial sarcoma	
AI801505	NCI_CGAP_Gas4	stomach	signet ring cells	-	-	adenocarcinoma	
AluSx Antisense Transcripts Ending in Promoter							
BG328312	NIH_MGC_15	colon	cell line	-	-	adenocarcinoma	
R73744	Soares placenta Nb2HP	placenta	-	placenta obtained at birth (full term)	female	normal	
AA297006	Adipose tissue, white II	adipose tissue, white	-	adult	-	normal	
AluSx Unknown Transcripts							
AI114755	Human fetal liver cDNA library	liver	-	fetus	-	normal	
AA468196	NCI_CGAP_Pr2	prostate gland	preneoplastic	adult, 45 year old	male	proneoplasia	
BF770715	IT0026	epididymis	-	adult	-	carcinoma	
AA614163	NCI_CGAP_Pr2	prostate gland	preneoplastic	adult, 45 year old	male	proneoplasia	
BE063437	BT0277	#N/A	#N/A	#N/A	#N/A	#N/A	
AW961994	MAGE resequences, MAGG	colon	-	-	-	tumor metastasis	
AI444575	Human fetal liver cDNA library	liver	-	fetus	-	normal	

BF887154	TN0174	#N/A	#N/A	#N/A	#N/A	#N/A	
AluY Sense Transcripts Starting in Promoter							
BG432302	NIH_MGC_75	kidney	-	-	-	normal	
BF964720	NIH_MGC_81	skeletal muscle	-	-	-	normal	
BF681576	NIH_MGC_83	prostate gland	-	-	male	normal	
BE297262	NIH_MGC_17	muscle	cell line	-	-	rhabdomyosarcoma	
BE389903	NIH_MGC_44	uterus (endometrium)	adenocarcinoma cell line	-	female	adenocarcinoma	
BG024252	NIH_MGC_88	small intestine	cell line	-	-	duodenal adenocarcinoma	
BE893169	NIH_MGC_72	skin	cell line	-	-	melanotic melanoma	
BF382973	NIH_MGC_56	brain	primitive neuroectoderm	-	-	normal	
AV764530	MDS	bone marrow	-	-	-	normal	
AW500226	NIH_MGC_50	lymph node	germinal centre B cells	-	-	normal	
AluY Sense Transcripts Ending in Promoter							
AA713815	NCI_CGAP_SS1	connective tissue	-	-	-	synovial sarcoma	
AW873530	Soares_NFL_T_GBC_S1	pooled (testis/lung/B-cell)	-	fetus	-	normal	
AA908422	NCI_CGAP_Ov8	ovary	-	-	female	serous adenocarcinoma	
BF942454	NCI_CGAP_Thy10	thyroid	-	-	-	medullary carcinoma	
AW050498	NCI_CGAP_Ut4	uterus	-	-	female	adenocarcinoma	
AW073470	NCI_CGAP_HSC2	bone marrow	CD34+/CD38- stem cells	adult	-	normal	
AA873532	NCI_CGAP_Kid5	kidney	clear cells	-	-	tumor (2 pooled)	
AA825357	NCI_CGAP_Lu5	lung	carcinoid	-	-	carcinoma	
AA985455	NCI_CGAP_Kid6	kidney	-	-	-	carcinoma	
AA346458	Fetal heart II	heart	-	fetus	-	normal	

AluY Antisense Transcripts Starting in Promoter							
AI349850	NCI_CGAP_Lu26	lung	-	adult	-	invasive adenocarcinoma	
AluY Antisense Transcripts Ending in Promoter							
BG027093	NIH_MGC_86	bone	cell line	-	-	osteosarcoma	
BI086493	NIH_MGC_10	cervix	MGC36	-	female	choriocarcinoma	
AL119691	761 (synonym: hamy2)	brain (amygdala)	-	adult	-	normal	
BG495281	NIH_MGC_59	salivary gland	cell line	-	-	mucoepidermoid carcinoma	
AL135405	762 (synonym: hmel2)	skin	cell line	adult	-	melanoma	
AA352803	Activated T-cells XX	lymphatic system	T-cells, activated	adult	-	normal	
AL596618	451 (synonym: hlcc1) spinal cord	spinal cord	-	adult	-	normal	
AL038705	566 (synonym: hfkd2)	kidney	-	fetus	-	normal	
H00934	Soares placenta Nb2HP	placenta	-	placenta obtained at birth (full term)	female	normal	
BG541823	NIH_MGC_77	lung	-	-	-	normal	
AluY Unknown Transcripts							
AA569205	NCI_CGAP_Lip2	adipose tissue	-	-	-	liposarcoma	
AA649705	NCI_CGAP_Alv1	lung	-	-	-	alveolar rhabdomyosarcoma	
AA542991	NCI_CGAP_Ov2	ovary	-	-	female	adenocarcinoma	
C05755	Human pancreatic islet	pancreas	pancreatic islets	adult	-	normal	
F17891	HM1	skeletal muscle (pectoral)	-	-	female	after mastectomy	
AA533725	NCI_CGAP_Pr11	prostate gland	prostatic epithelial cells	-	male	normal	
AW961889	MAGE resequences, MAGG	colon	-	-	-	tumor metastasis	
BF902602	MT0190	bone marrow	-	adult	-	chronic myelogenous leukemia	
AA613345	NCI_CGAP_Thy1	thyroid	-	-	-	carcinoma	
M86143	Hippocampus, Stratagene (cat. #936205)	brain (hippocampus)	-	2 years old	female	normal	

AluYb8 Sense Transcripts Starting in Promoter							
BE390977	NIH_MGC_44	uterus (endometrium)	adenocarcinoma cell line	-	female	adenocarcinoma	
BE277602	NIH_MGC_20	skin	cell line	-	-	melanotic melanoma	
BF678329	NIH_MGC_83	prostate gland	-	-	male	normal	
BG036934	NIH_MGC_96	brain (hypothalamus)	-	-	-	normal	
AV761056	MDS	bone marrow	-	-	-	normal	
BG541138	NIH_MGC_77	lung	-	-	-	normal	
BE541040	NIH_MGC_10	cervix	MGC36	-	female	choriocarcinoma	
BF248091	NIH_MGC_58	kidney	cell line	-	-	hypernephroma	
BG531704	NIH_MGC_61	testis	cell line	embryo	male	embryonal carcinoma	
BF965180	NIH_MGC_81	skeletal muscle	-	-	-	normal	
AluYb8 Sense Transcripts Ending in Promoter							
AI918421	NCI_CGAP_Pan1	pancreas	-	-	-	adenocarcinoma	
AW088202	NCI_CGAP_Brn35	brain	pooled	-	-	astrocytoma/ glioblastoma multiforme /meningioma/oligodendroglioma/medulloblastoma	
AA634146	Stratagene hNT neuron (#937233)	nervous tissue	hNT neurons (differentiated, post mitotic)	-	-	normal	
AA856858	NCI_CGAP_Ov2	ovary	-	-	female	adenocarcinoma	
H78195	Soares fetal liver spleen 1NFLS	liver and spleen	-	fetus, 20 week post conception	male	normal	
AA714055	NCI_CGAP_GCB0	tonsil	B-cells, germinal centre	-	-	normal	
AW079659	NCI_CGAP_Eso2	esophagus	squamous cells	-	-	squamous cell carcinoma	
AW075511	NCI_CGAP_Kid13	kidney	-	-	-	Wilms' tumors	
AI568678	NCI_CGAP_CLL1	lymphatic system	B-cells	-	-	chronic lymphotic leukemia	
N63352	Morton Fetal Cochlea	ear (cochlea)	-	fetus, 16-22 weeks	-	normal	

AluYb8 Antisense Transcripts Starting in Promoter							
AW419118	NCI_CGAP_Lu34.1	#N/A	#N/A	#N/A	#N/A	#N/A	
AI633007	NCI_CGAP_Ut2	uterus	-	-	female	adenocarcinoma	
AluYb8 Antisense Transcripts Ending in Promoter							
AA488573	Stratagene HeLa cell s3 937216	cervix	HeLa S3 cells	-	female	carcinoma	
AA132833	Stratagene colon (#937204)	colon	T84 carcinoma cell line	-	-	carcinoma	
AV764179	MDS	bone marrow	-	-	-	normal	
BI825211	NIH_MGC_119	brain (medulla)	-	-	-	normal	
AA021367	Soares retina N2b4HR	eye (retina)	-	adult, 55 year old	male	normal	
R47855	Soares breast 2NbHBst	mammary gland	-	adult	female	normal	
AL599804	313 (synonym: hlcc2)	uncharacterized tissue	-	adult	-	-	
H50727	Soares fetal liver spleen 1NFLS	liver and spleen	-	fetus, 20 week post conception	male	normal	
AV732919	cdA	adrenal gland	-	adult	-	pheochromocytoma	
BI464563	NIH_MGC_97	testis	-	-	male	normal	
AluYb8 Unknown Transcripts							
AI064952	Human fetal liver cDNA library	liver	-	fetus	-	normal	
AA837084	NCI_CGAP_GCB1	tonsil	B-cells, germinal centre	-	-	normal	
AA523815	NCI_CGAP_Ov2	ovary	-	-	female	adenocarcinoma	
C06458	Human pancreatic islet	pancreas	pancreatic islets	adult	-	normal	
AA729721	NCI_CGAP_Alv1	lung	-	-	-	alveolar rhabdomyosarcoma	
AA550758	NCI_CGAP_Pr11	prostate gland	prostatic epithelial cells	-	male	normal	
F17761	HM1	skeletal muscle (pectoral)	-	-	female	after mastectomy	
AA480790	NCI_CGAP_Kid1	kidney	renal cells	-	-	renal cell carcinoma	
BF829536	HN0034	thyroid	-	adult	-	normal	
BF896102	MT0157	blood	-	adult	-	chronic myelogenous leukemia	

AluYd2 Sense Transcripts Starting in Promoter							
BG435460	NIH_MGC_79	placenta	-	-	-	normal	
AV763007	MDS	bone marrow	-	-	-	normal	
BI850201	NIH_MGC_83	prostate gland	-	-	male	normal	
BG429918	NIH_MGC_75	kidney	-	-	-	normal	
BF665455	NIH_MGC_56	brain	primitive neuroectoderm	-	-	normal	
BG576979	NIH_MGC_87	mammary gland	cell line	-	female	mammary adenocarcinoma	
T58118	Stratagene fetal spleen (#937205)	spleen	-	fetus	-	normal	
BG115611	NIH_MGC_88	small intestine	cell line	-	-	duodenal adenocarcinoma	
BF965231	NIH_MGC_81	skeletal muscle	-	-	-	normal	
BG777070	NIH_MGC_59	salivary gland	cell line	-	-	mucoepidermoid carcinoma	
AluYd2 Sense Transcripts Ending in Promoter							
BM353613	HR85 islet	pancreas	pancreatic islet	-	-	normal	
AA578861	NCI_CGAP_Sch1	nervous tissue	-	-	-	Schwannoma tumour	
AI241821	NCI_CGAP_Brn35	brain	pooled	-	-	astrocytoma/ glioblastoma multiforme /meningioma/oligodendroglioma/medulloblastoma	
AI783494	NCI_CGAP_Ut2	uterus	-	-	female	adenocarcinoma	
BE045001	NCI_CGAP_Thy7	thyroid	-	-	-	follicular adenoma (benign lesion)	
AW082492	NCI_CGAP_Eso2	esophagus	squamous cells	-	-	squamous cell carcinoma	
AI886903	NCI_CGAP_Ut1	uterus	-	-	female	adenocarcinoma	
AA937037	NCI_CGAP_Kid3	kidney	-	-	-	normal	
AA812281	NCI_CGAP_Pr24	prostate gland	cell line	-	male	invasive tumor (adenocarcinoma)	
AA788982	Soares_parathyroid_tumor_NbHPA	parathyroid gland	-	adult	-	adenoma	
AluYd2 Antisense Transcripts Starting in Promoter							
AI583283	NCI_CGAP_HSC4	#N/A	#N/A	#N/A	#N/A	#N/A	
AI872020	NCI_CGAP_Ut2	uterus	-	-	female	adenocarcinoma	

AA055169	Soares_fetal_heart_NbHH19W	heart	-	fetus, 19 week post conception	-	normal	
AluYd2 Antisense Transcripts Ending in Promoter							
AL036382	564 (synonym: hfbr2)	brain	-	fetus	-	normal	
AA300061	Uterus tumor I	uterus	-	adult	female	tumor (neoplasia)	
BG181119	NIH_MGC_91	prostate gland	cell line	-	male	adenocarcinoma	
AL603093	686 (synonym: hlcc3)	skeletal muscle	-	adult	-	normal	
AA634889	Stratagene lung (#937210)	lung	-	adult, 72 year old	male	normal	
R86151	Soares fetal liver spleen 1NFLS	liver and spleen	-	fetus, 20 week post conception	male	normal	
R24887	Soares placenta Nb2HP	placenta	-	placenta obtained at birth (full term)	female	normal	
N94233	Soares fetal liver spleen 1NFLS	liver and spleen	-	fetus, 20 week post conception	male	normal	
H86431	Soares retina N2b5HR	eye (retina)	-	adult, 55 year old	male	normal	
BI755034	NIH_MGC_114	brain	-	-	-	normal	
AluYd2 Unknown Transcripts							
AW972628	MAGE resequences, MAGL	colon	-	-	-	tumor metastasis	
AA525898	NCI_CGAP_Ov2	ovary	-	-	female	adenocarcinoma	
AL548579	Homo sapiens PLACENTA COT 25-NORMALIZED	placenta	-	-	female	normal	
AA177120	NCI_CGAP_Pr2	prostate gland	preneoplastic	adult, 45 year old	male	proneoplasia	
AA502860	NCI_CGAP_Lip2	adipose tissue	-	-	-	liposarcoma	
AW801242	UM0065	uterus	-	adult	female	normal	
F17142	HM1	skeletal muscle (pectoral)	-	-	female	after mastectomy	

AA632529	NCI_CGAP_Thy1	thyroid	-	-	-	carcinoma	
AA492081	NCI_CGAP_Kid1	kidney	renal cells	-	-	renal cell carcinoma	
AI133297	Human fetal liver cDNA library	liver	-	fetus	-	normal	

IN SILICO MOUSE RETROELEMENT TRANSCRIPTS

ACCESSION NUMBER	CLONE LIBRARY	TISSUE	CELL TYPE	DEV. STAGE	SEX	DISEASE STATE
ETn Sense Transcripts Starting in LTR						
AA097639	Life Tech mouse embryo 10 5dpc 10665016	embryo	-	embryo, 10.5 day post conception	-	normal
BE850132	Soares mouse 3NbMS	spleen	-	4 weeks	male	normal
AA171078	Life Tech mouse embryo 13 5dpc 10666014	embryo	-	embryo, 13.5 day post conception	-	normal
BF159452	NCI_CGAP_Lu29	lung	-	-	-	-
AW215348	NCI_CGAP_Lu30	lung	-	-	-	tumor, metastatic to mammary
BE627087	Soares_mammary_gland_NMLMG	mammary gland	-	adult	-	normal
BB594285	RIKEN full-length enriched, 4 days neonate thymus	thymus	-	neonate, 4 day	-	normal
ETn Sense Transcripts Ending in LTR						
AW212422	NCI_CGAP_Mam3	#N/A	#N/A	#N/A	#N/A	#N/A
BF147344	Soares_NMEBA_branchial_arch	embryo	branchial arches	embryo, 10.5 day post conception	-	normal
AI891899	Sugano mouse kidney mkia	kidney	-	adult	female	normal
AI891530	Sugano mouse kidney mkia	kidney	-	adult	female	normal
AI255243	Sugano mouse liver mlia	liver	-	adult	female	normal
BF319817	NCI_CGAP_Mam5	mammary gland	gross tissue	7 months	female	tumor
AI662415	Stratagene mouse skin (#937313)	skin (whole)	-	11 weeks	female	normal
AI425880	Soares mouse embryo NbME13.5 14.5	embryo	-	embryo, 13.5-14.5 day post conception total f	-	-
C78605	Mouse 3.5-dpc blastocyst cDNA	-	-	blastocyst, 3.5 day post conception	-	normal
AI853958	NIH_BMAP_M_S1	brain	-	27-32 days	-	normal

ETn Antisense Transcripts Starting in LTR						
AA089257	Stratagene mouse Tcell 937311	lymphatic system	T-cells (M30 CD4+ cells)	-	-	normal
AA726967	Stratagene mouse Tcell 937311	lymphatic system	T-cells (M30 CD4+ cells)	-	-	normal
AA065614	Stratagene mouse kidney (#937315)	kidney	-	4 weeks	-	normal
AA760040	Stratagene mouse skin (#937313)	skin (whole)	-	11 weeks	female	normal
AA794278	Stratagene mouse skin (#937313)	skin (whole)	-	11 weeks	female	normal
W71327	Soares mouse embryo NbME13.5 14.5	embryo	-	embryo, 13.5-14.5 day post conception total f	-	-
AA388210	Ko mouse embryo 11 5dpc	embryo	-	embryo, 11.5 day post conception	-	normal
BG081454	NIA Mouse 15K cDNA Clone Set	embryo (pooled)	-	pooled	-	normal
AA589016	Beddington mouse embryonic region	embryo	-	embryo, 7.5 day post conception	-	normal
BF166031	NCI_CGAP_Lu29	lung	-	-	-	-
ETn Unknown Direction						
C89169	Mouse early blastocyst cDNA	-	-	blastocyst	-	normal
GLN Sense Transcripts Starting in LTR						
BB840929	RIKEN full-length enriched, 6 days neonate spleen	spleen	-	neonate, 6 day	-	normal
BF731574	NCI_CGAP_Sp2	spleen	NK cells (flow-sorted)	-	-	-
BF165282	NCI_CGAP_Lu29	lung	-	-	-	-
BF160675	NCI_CGAP_Lu29	lung	-	-	-	-
AW215122	NCI_CGAP_Lu30	lung	-	-	-	tumor, metastatic to mammary
BI555829	NCI_CGAP_Mam3	#N/A	#N/A	#N/A	#N/A	#N/A
BI149358	NCI_CGAP_Lu29	lung	-	-	-	-
AW762831	NCI_CGAP_Mam3	#N/A	#N/A	#N/A	#N/A	#N/A
BI151436	NCI_CGAP_Lu29	lung	-	-	-	-
AW323376	NCI_CGAP_Lu29	lung	-	-	-	-

GLN Sense Transcripts Ending in LTR						
BB670907	RIKEN full-length enriched, 10 days embryo	embryo	-	embryo, 10 days	-	normal
BB800634	RIKEN full-length enriched, adult male accessory axillary lymph node	lymph node (accessory axillary)	-	male	-	-
AV303415	RIKEN full-length enriched, 8 days embryo	embryo	-	embryo, 8 day post conception	-	normal
BB671022	RIKEN full-length enriched, 10 days embryo	embryo	-	embryo, 10 days	-	normal
BB793888	RIKEN full-length enriched, RCB-0035 WEHI-3 cDNA	-	-	-	-	-
BB802309	RIKEN full-length enriched, adult male accessory axillary lymph node	lymph node (accessory axillary)	-	male	-	-
BB670724	RIKEN full-length enriched, 10 days embryo	embryo	-	embryo, 10 days	-	normal
BB755859	RIKEN full-length enriched, melanocyte	skin	melanocyte	-	-	normal
BB727075	RIKEN full-length enriched, 8 cells embryo	-	-	embryo (8 cell)	-	-
BB790475	RIKEN full-length enriched, B cells CRL-1669 BCL1 Clone 13.20-3B3 cDNA	-	CRL-1669 BCL1 Clone 13.20-3B3 B-cells	-	-	leukemia
GLN Antisense Transcripts Starting in LTR						
AI504983	Stratagene mouse diaphragm (#937303)	diaphragm	-	adult	-	normal
AI451942	Soares_thymus_2NbMT	thymus	-	4 weeks	male	normal
AI606260	Stratagene mouse skin (#937313)	skin (whole)	-	11 weeks	female	normal
GLN Antisense Transcripts Ending in LTR						
BI686827	NCI_CGAP_Mam6	mammary gland	-	5 months	-	infiltrating ductal carcinoma
BB663321	RIKEN full-length enriched, 15 days embryo head	head	-	embryo, 15 day post conception	-	normal
BG089161	Soares_mouse_NMGB_bcell	lymphatic system	B-cells, germinal	-	-	normal
BE650076	NIH_BMAP_M_S4	brain	-	27-32 days	-	normal

GLN Unknown Direction						
AV286981	RIKEN full-length enriched, 11 days pregnant adult female ovary and uterus	ovary and uterus	-	adult, 11 days pregnant	female	normal
AV024160	Mus musculus adult C57BL/6J lung	lung	-	adult	-	normal
AV173083	Mus musculus liver C57BL/6J 14-day embryo	#N/A	#N/A	#N/A	#N/A	#N/A
AV119028	Mus musculus C57BL/6J 10-day embryo	-	-	embryo, 10 day post conception	-	normal
AV111767	Mus musculus C57BL/6J 10-day embryo	-	-	embryo, 10 day post conception	-	normal
MeRV-L Sense Transcripts Starting in LTR						
AA437897	Knowles Solter mouse 2 cell	embryo	-	embryo, 2-cell stage	-	normal
AA549714	Knowles Solter mouse 2 cell	embryo	-	embryo, 2-cell stage	-	normal
AA692869	Knowles Solter mouse 2 cell	embryo	-	embryo, 2-cell stage	-	normal
AA438239	Knowles Solter mouse 2 cell	embryo	-	embryo, 2-cell stage	-	normal
AA792022	Knowles Solter mouse 2 cell	embryo	-	embryo, 2-cell stage	-	normal
AA647938	Knowles Solter mouse 2 cell	embryo	-	embryo, 2-cell stage	-	normal
AA692722	Knowles Solter mouse 2 cell	embryo	-	embryo, 2-cell stage	-	normal
AA636442	Knowles Solter mouse 2 cell	embryo	-	embryo, 2-cell stage	-	normal
AA549375	Knowles Solter mouse 2 cell	embryo	-	embryo, 2-cell stage	-	normal
AA547414	Knowles Solter mouse 2 cell	embryo	-	embryo, 2-cell stage	-	normal
MeRV-L Sense Transcripts Ending in LTR						
AA415479	Knowles Solter mouse 2 cell	embryo	-	embryo, 2-cell stage	-	normal
AA415429	Knowles Solter mouse 2 cell	embryo	-	embryo, 2-cell stage	-	normal
AA415049	Knowles Solter mouse 2 cell	embryo	-	embryo, 2-cell stage	-	normal
AA673473	Knowles Solter mouse 2 cell	embryo	-	embryo, 2-cell stage	-	normal
AA792903	Knowles Solter mouse 2 cell	embryo	-	embryo, 2-cell stage	-	normal
AA419630	Knowles Solter mouse 2 cell	embryo	-	embryo, 2-cell stage	-	normal
AA816119	Knowles Solter mouse 2 cell	embryo	-	embryo, 2-cell stage	-	normal
AA792882	Knowles Solter mouse 2 cell	embryo	-	embryo, 2-cell stage	-	normal
AA672927	Knowles Solter mouse 2 cell	embryo	-	embryo, 2-cell stage	-	normal
AA672938	Knowles Solter mouse 2 cell	embryo	-	embryo, 2-cell stage	-	normal

MeRV-L Antisense Transcripts Starting in LTR						
AA624122	Knowles Solter mouse 2 cell	embryo	-	embryo, 2-cell stage	-	normal
AA681111	Knowles Solter mouse 2 cell	embryo	-	embryo, 2-cell stage	-	normal
MeRV-L Antisense Transcripts Ending in LTR						
BB289195	RIKEN full-length enriched, 2 cells egg	egg	-	embryo, 2-cell stage	-	normal
AI505665	Knowles Solter mouse 2 cell	embryo	-	embryo, 2-cell stage	-	normal
AA939921	Knowles Solter mouse 2 cell	embryo	-	embryo, 2-cell stage	-	normal
AI639935	Knowles Solter mouse 2 cell	embryo	-	embryo, 2-cell stage	-	normal
AI505676	Knowles Solter mouse 2 cell	embryo	-	embryo, 2-cell stage	-	normal
AA675339	Knowles Solter mouse 2 cell	embryo	-	embryo, 2-cell stage	-	normal
AI505206	Knowles Solter mouse 2 cell	embryo	-	embryo, 2-cell stage	-	normal
AI505950	Stratagene mouse Tcell 937311	lymphatic system	T-cells (M30 CD4+ cells)	-	-	normal
AI586307	Stratagene mouse Tcell 937311	lymphatic system	T-cells (M30 CD4+ cells)	-	-	normal
AI561808	Stratagene mouse skin (#937313)	skin (whole)	-	11 weeks	female	normal
MeRV-L Unknown Direction						
AJ133289	Mus musculus NEURO-2a	#N/A	#N/A	#N/A	#N/A	#N/A
AJ133283	Mus musculus NEURO-2a	#N/A	#N/A	#N/A	#N/A	#N/A
MMTV Sense Transcripts Starting in LTR						
BI456557	NCI_CGAP_Mam5	mammary gland	gross tissue	7 months	female	tumor
BI454696	NCI_CGAP_Mam5	mammary gland	gross tissue	7 months	female	tumor
BG173972	NCI_CGAP_Mam1	mammary gland	biopsy sample	3 months, virgin	female	tumor
BI652911	NCI_CGAP_Mam3	#N/A	#N/A	#N/A	#N/A	#N/A
BI452662	NCI_CGAP_Mam5	mammary gland	gross tissue	7 months	female	tumor
BI414622	NCI_CGAP_Lu33	lung	-	-	-	lung tumor
BG247571	NCI_CGAP_Mam1	mammary gland	biopsy sample	3 months, virgin	female	tumor
BI109657	NCI_CGAP_Mam5	mammary gland	gross tissue	7 months	female	tumor
BG175359	NCI_CGAP_Mam1	mammary gland	biopsy sample	3 months, virgin	female	tumor
BE381888	NCI_CGAP_Mam1	mammary gland	biopsy sample	3 months, virgin	female	tumor

MMTV Sense Transcripts Ending in LTR						
BB792747	RIKEN full-length enriched, RCB-0035 WEHI-3 cDNA	-	-	-	-	-
BB109769	RIKEN full-length enriched, adult male urinary bladder	bladder	-	adult	male	normal
BB225007	RIKEN full-length enriched, adult male aorta and vein	aorta and vein	-	adult	male	normal
BB786806	RIKEN full-length enriched, B cells CRL-1669 BCL1 Clone 13.20-3B3 cDNA	-	CRL-1669 BCL1 Clone 13.20-3B3 B-cells	-	-	leukemia
BB788763	RIKEN full-length enriched, B cells CRL-1669 BCL1 Clone 13.20-3B3 cDNA	-	CRL-1669 BCL1 Clone 13.20-3B3 B-cells	-	-	leukemia
BB784547	RIKEN full-length enriched, B cells CRL-1669 BCL1 Clone 13.20-3B3 cDNA	-	CRL-1669 BCL1 Clone 13.20-3B3 B-cells	-	-	leukemia
BB780456	RIKEN full-length enriched, B cells CRL-1669 BCL1 Clone 13.20-3B3 cDNA	-	CRL-1669 BCL1 Clone 13.20-3B3 B-cells	-	-	leukemia
BB791184	RIKEN full-length enriched, B cells CRL-1669 BCL1 Clone 13.20-3B3 cDNA	-	CRL-1669 BCL1 Clone 13.20-3B3 B-cells	-	-	leukemia
BB789527	RIKEN full-length enriched, B cells CRL-1669 BCL1 Clone 13.20-3B3 cDNA	-	CRL-1669 BCL1 Clone 13.20-3B3 B-cells	-	-	leukemia
BB784531	RIKEN full-length enriched, B cells CRL-1669 BCL1 Clone 13.20-3B3 cDNA	-	CRL-1669 BCL1 Clone 13.20-3B3 B-cells	-	-	leukemia
MMTV Antisense Transcripts Starting in LTR						
BB509585	RIKEN full-length enriched, 10 days lactation, adult female mammary gland	mammary gland	-	adult, 10 days lactation	female	normal
BB505838	RIKEN full-length enriched, 10 days lactation, adult female mammary gland	mammary gland	-	adult, 10 days lactation	female	normal

BB512893	RIKEN full-length enriched, 10 days lactation, adult female mammary gland	mammary gland	-	adult, 10 days lactation	female	normal
BB507012	RIKEN full-length enriched, 10 days lactation, adult female mammary gland	mammary gland	-	adult, 10 days lactation	female	normal
BB512732	RIKEN full-length enriched, 10 days lactation, adult female mammary gland	mammary gland	-	adult, 10 days lactation	female	normal
MMTV Antisense Transcripts Ending in LTR						
BE303171	NCI_CGAP_Lu29	lung	-	-	-	-
BI695121	NCI_CGAP_Mam2	mammary gland	biopsy sample	5 months	female	tumor
BI452499	NCI_CGAP_Mam5	mammary gland	gross tissue	7 months	female	tumor
BI157243	NCI_CGAP_Mam3	#N/A	#N/A	#N/A	#N/A	#N/A
MuLV Sense Transcripts Starting in LTR						
BG975178	NCI_CGAP_Mam4	#N/A	#N/A	#N/A	#N/A	#N/A
AL363277	ICRFp 522 and 523	embryo	-	embryo, 9 and 12 day post conception	-	normal
BG915150	NCI_CGAP_Mam4	#N/A	#N/A	#N/A	#N/A	#N/A
BG862909	NCI_CGAP_Mam4	#N/A	#N/A	#N/A	#N/A	#N/A
BF384627	NCI_CGAP_Li9	liver	-	-	-	normal
BG974033	NCI_CGAP_Mam4	#N/A	#N/A	#N/A	#N/A	#N/A
BG975178	NCI_CGAP_Mam4	#N/A	#N/A	#N/A	#N/A	#N/A
BF137408	NCI_CGAP_Lu30	lung	-	-	-	tumor, metastatic to mammary
BG915357	NCI_CGAP_Mam4	#N/A	#N/A	#N/A	#N/A	#N/A
BF235889	NCI_CGAP_Li9	liver	-	-	-	normal

MuLV Sense Transcripts Ending in LTR						
BE951976	NIH_BMAP_Ret1	eye (retina)	-	embryo, 13.5 days post-conception	-	normal
BG073049	NIA Mouse 15K cDNA Clone Set	embryo (pooled)	-	pooled	-	normal
AU041330	Mouse four-cell-embryo cDNA	-	-	embryo, four-cell	-	normal
AI839695	NIH_BMAP_MBG	brain (basal ganglia)	-	27-32 days	-	normal
AW111317	mouse liver, dioxin treated	liver	-	-	male	normal
AU043054	Mouse four-cell-embryo cDNA	-	-	embryo, four-cell	-	normal
AU042915	Mouse four-cell-embryo cDNA	-	-	embryo, four-cell	-	normal
AU042635	Mouse four-cell-embryo cDNA	-	-	embryo, four-cell	-	normal
AU042455	Mouse four-cell-embryo cDNA	-	-	embryo, four-cell	-	normal
AU042235	Mouse four-cell-embryo cDNA	-	-	embryo, four-cell	-	normal
MuLV Antisense Transcripts Starting in LTR						
BF533001	NCI_CGAP_Li9	liver	-	-	-	normal
MuLV Antisense Transcripts Ending in LTR						
AV251543	RIKEN full-length enriched, 0 day neonate head	head	-	neonate, 0 day	-	-
MuLV Unknown Direction						
AV077396	Mus musculus stomach C57BL/6J adult	stomach	-	adult	male	normal
VL30 Sense Transcripts Starting in LTR						
AI391174	Soares mouse p3NMF19.5	-	-	fetus, 19.5 day post conception	-	normal
W83035	Soares mouse p3NMF19.5	-	-	fetus, 19.5 day post conception	-	normal
AW318911	Sugano mouse kidney mkia	kidney	-	adult	female	normal
BF181609	NCI_CGAP_Mam5	mammary gland	gross tissue	7 months	female	tumor
BF179435	NCI_CGAP_Mam5	mammary gland	gross tissue	7 months	female	tumor
BB603779	RIKEN full-length enriched, 15 days embryo head	head	-	embryo, 15 day post conception	-	normal
AA462750	Soares_mammary_gland_NbMMG	mammary gland	-	4 weeks	male	normal

BB569651	RIKEN full-length enriched, 13 days embryo	-	-	embryo, 13 day post conception	-	normal
BB574170	RIKEN full-length enriched, 11 days pregnant adult female ovary and uterus	ovary and uterus	-	adult, 11 days pregnant	female	normal
BB570887	RIKEN full-length enriched, 0 day neonate skin	skin	-	neonate, 0 day	-	normal
VL30 Sense Transcripts Ending in LTR						
AI506790	Stratagene mouse Tcell 937311	lymphatic system	T-cells (M30 CD4+ cells)	-	-	normal
BE456905	Soares_NMEBA_branchial_arch	embryo	branchial arches	embryo, 10.5 day post conception	-	normal
BF719425	Soares_NMEBA_branchial_arch	embryo	branchial arches	embryo, 10.5 day post conception	-	normal
AW822120	Ren Stubbs mouse thymus	thymus	-	3 weeks	-	normal
AI788547	Sugano mouse kidney mkia	kidney	-	adult	female	normal
AI256279	Sugano mouse liver mlia	liver	-	adult	female	normal
AI195472	Sugano mouse liver mlia	liver	-	adult	female	normal
AI786275	Sugano mouse liver mlia	liver	-	adult	female	normal
AW240981	NCI_CGAP_Mam5	mammary gland	gross tissue	7 months	female	tumor
BF227389	NCI_CGAP_Mam5	mammary gland	gross tissue	7 months	female	tumor
VL30 Antisense Transcripts Starting in LTR						
AA545096	Knowles Solter mouse 2 cell	embryo	-	embryo, 2-cell stage	-	normal
AA086644	Stratagene mouse Tcell 937311	lymphatic system	T-cells (M30 CD4+ cells)	-	-	normal
AA546016	Stratagene mouse Tcell 937311	lymphatic system	T-cells (M30 CD4+ cells)	-	-	normal
AA656013	Stratagene mouse Tcell 937311	lymphatic system	T-cells (M30 CD4+ cells)	-	-	normal
AA646267	Stratagene mouse Tcell 937311	lymphatic system	T-cells (M30 CD4+ cells)	-	-	normal
AA556012	Stratagene mouse Tcell 937311	lymphatic system	T-cells (M30 CD4+ cells)	-	-	normal
AA109808	Stratagene mouse kidney (#937315)	kidney	-	4 weeks	-	normal

AA727665	Stratagene mouse skin (#937313)	skin (whole)	-	11 weeks	female	normal
AA798773	Stratagene mouse skin (#937313)	skin (whole)	-	11 weeks	female	normal
AA530160	Stratagene mouse skin (#937313)	skin (whole)	-	11 weeks	female	normal
VL30 Antisense Transcripts Ending in LTR						
AI644589	Stratagene mouse heart (#937316)	heart	-	embryo, 13 day post conception	-	normal
L1A Sense Transcripts Starting in Promoter						
AA718056	Soares_mammary_gland_NbMMG	mammary gland	-	4 weeks	male	normal
BB623589	RIKEN full-length enriched mouse cDNA library, C57BL/6J mullerian duct includes surrounding region female 12 days embryo	mullerian duct includes surrounding region	-	12 days embryo	female	-
AA269437	Soares mouse NML	liver	-	adult	-	normal
L1A Antisense Transcripts Starting in Promoter						
BE954880	NIH_BMAP_Ret2_N	eye (retina)	-	neonate, 1 day	-	normal
BI133574	NIH_BMAP_M_S4	brain	-	27-32 days	-	normal
BI791025	Melton Normalized Mixed Mouse Pancreas 1 N1-MMS1	pancreatic islet	-	Embryonic day 10.5, E12.5, E16.5, newborn, ad	-	normal
AU015217	Mouse two-cell stage embryo cDNA	-	-	embryo, 2-cell stage	-	normal
L1A Antisense Transcripts Ending in Promoter						
BG078622	NIA Mouse 15K cDNA Clone Set	embryo (pooled)	-	pooled	-	normal
BG083145	NIA Mouse 15K cDNA Clone Set	embryo (pooled)	-	pooled	-	normal
AA098176	Stratagene mouse Tcell 937311	lymphatic system	T-cells (M30 CD4+ cells)	-	-	normal
BB664412	RIKEN full-length enriched, 0 day neonate lung	lung	-	neonate, 0 day	-	normal
BI791329	Melton Normalized Mixed Mouse Pancreas 1 N1-MMS1	pancreatic islet	-	Embryonic day 10.5, E12.5, E16.5, newborn, ad	-	normal
BB651716	RIKEN full-length enriched, ES cells	-	embryonic stem cells	embryo	-	normal

AA656443	Stratagene mouse skin (#937313)	skin (whole)	-	11 weeks	female	normal
BG088687	NIA Mouse 15K cDNA Clone Set	embryo (pooled)	-	pooled	-	normal
BB650488	RIKEN full-length enriched, 0 day neonate cerebellum	brain (cerebellum)	-	neonate, 0 day	-	normal
AA791912	Stratagene mouse skin (#937313)	skin (whole)	-	11 weeks	female	normal
L1A Unknown Direction						
BI986120	Mouse E14.5 retina lambda ZAP II Library	eye	neural retina	embryonic day 14.5 post-fertilization	-	normal
BG803570	Mouse E14.5 retina lambda ZAP II Library	eye	neural retina	embryonic day 14.5 post-fertilization	-	normal
AV172158	Mus musculus head C57BL/6J 17-day embryo	head	-	embryo, 17 day post conception	-	normal
BI986243	Mouse E14.5 retina lambda ZAP II Library	eye	neural retina	embryonic day 14.5 post-fertilization	-	normal
L1Gf Sense Transcripts Starting in Promoter						
BB646138	RIKEN full-length enriched, 10 days neonate cerebellum	brain (cerebellum)	-	neonate, 10 day	-	normal
BB648069	RIKEN full-length enriched, 16 days embryo head	head	-	embryo, 16 day post conception	-	normal
BB616730	RIKEN full-length enriched, adult male testis	testis	-	adult	male	normal
BB616742	RIKEN full-length enriched, adult male testis	testis	-	adult	male	normal
BG242892	NCI_CGAP_Mam1	mammary gland	biopsy sample	3 months, virgin	female	tumor
BB616213	RIKEN full-length enriched mouse cDNA library, C57BL/6J testis male adult	testis	-	adult	male	-
BB616883	RIKEN full-length enriched mouse cDNA library, C57BL/6J testis male adult	testis	-	adult	male	-
BB616381	RIKEN full-length enriched mouse cDNA library, C57BL/6J testis male adult	testis	-	adult	male	-
BF714790	Soares_thymus_2NbMT	thymus	-	4 weeks	male	normal

AA717102	Stratagene mouse skin (#937313)	skin (whole)	-	11 weeks	female	normal
L1Gf Sense Transcripts Ending in Promoter						
BB808667	RIKEN full-length enriched, 11 days embryo spinal cord	spinal cord	-	embryo, 11 day post conception	-	normal
BB395901	RIKEN full-length enriched, 0 day neonate cerebellum	brain (cerebellum)	-	neonate, 0 day	-	normal
BB726217	RIKEN full-length enriched, 8 cells embryo	-	-	embryo (8 cell)	-	-
BB052788	RIKEN full-length enriched, 12 days embryo male wolffian duct	-	wolffian duct includes surrounding region	embryo, 12 day post conception	male	normal
AI449283	Soares mouse 3NbMS	spleen	-	4 weeks	male	normal
BB356908	RIKEN full-length enriched, adult male corpus striatum	brain (corpus striatum)	-	adult	male	-
BB156542	RIKEN full-length enriched, 16 days neonate thymus	thymus	-	neonate, 16 day	-	normal
AV241977	RIKEN full-length enriched, 0 day neonate head	head	-	neonate, 0 day	-	-
BB121843	RIKEN full-length enriched, adult male urinary bladder	bladder	-	adult	male	normal
BB318244	RIKEN full-length enriched, adult male corpora quadrigemina	brain (corpora quadrigemina)	-	adult	male	normal
L1Gf Antisense Transcripts Starting in Promoter						
AU024664	Mouse unfertilized egg cDNA	egg	-	egg, unfertilized	-	normal
AU021811	Mouse unfertilized egg cDNA	egg	-	egg, unfertilized	-	normal
AU021820	Mouse unfertilized egg cDNA	egg	-	egg, unfertilized	-	normal
C80423	Mouse 3.5-dpc blastocyst cDNA	-	-	blastocyst, 3.5 day post conception	-	normal
AW049672	NIH_BMAP_M_S2	brain	-	27-32 days	-	normal
BE995402	NIH_BMAP_Ret4_S2	eye (retina)	-	pooled	-	normal
C78246	Mouse 3.5-dpc blastocyst cDNA	-	-	blastocyst, 3.5 day post conception	-	normal
AU023403	Mouse unfertilized egg cDNA	egg	-	egg, unfertilized	-	normal
AU023408	Mouse unfertilized egg cDNA	egg	-	egg, unfertilized	-	normal

AU023460	Mouse unfertilized egg cDNA	egg	-	egg, unfertilized	-	normal
L1Gf Antisense Transcripts Ending in Promoter						
BI734500	NIH_MGC_94	eye (retina)	-	-	-	normal
AW475606	Sugano mouse kidney mkia	kidney	-	adult	female	normal
AA088077	Stratagene mouse lung 937302	lung	-	6-8 months	female	normal
AL362433	ICRFp 522 and 523	embryo	-	embryo, 9 and 12 day post conception	-	normal
BI466761	Kaestner ngn3 wt	pancreas	-	embryo, 14.5 day	-	normal
BI689256	NCI_CGAP_Mam6	mammary gland	-	5 months	-	infiltrating ductal carcinoma
BB839597	RIKEN full-length enriched, 8 cells embryo	-	-	embryo (8 cell)	-	-
BB838604	RIKEN full-length enriched, 8 cells embryo	-	-	embryo (8 cell)	-	-
AA114559	Beddington mouse embryonic region	embryo	-	embryo, 7.5 day post conception	-	normal
BE687208	Soares mouse 3NbMS	spleen	-	4 weeks	male	normal
L1Gf Unknown Direction						
BE192542	R3TA	adipose tissue	undifferentiated 3T3 cell line	-	-	-
BG800901	Mouse E14.5 retina lambda ZAP II Library	eye	neural retina	embryonic day 14.5 post-fertilization	-	normal
BI816982	Mouse prepuberal testis cDNA library	#N/A	#N/A	#N/A	#N/A	#N/A
BG806871	Mouse E14.5 retina lambda ZAP II Library	eye	neural retina	embryonic day 14.5 post-fertilization	-	normal
BI068108	Mouse embryonic 17.5 day pancreatic islet library	#N/A	#N/A	#N/A	#N/A	#N/A
BG801615	Mouse E14.5 retina lambda ZAP II Library	eye	neural retina	embryonic day 14.5 post-fertilization	-	normal
L1Tf Sense Transcripts Starting in Promoter						
AW320166	NCI_CGAP_Mam6	mammary gland	-	5 months	-	infiltrating ductal carcinoma
AA612097	Stratagene mouse skin (#937313)	skin (whole)	-	11 weeks	female	normal

BB644772	RIKEN full-length enriched, adult male corpora quadrigemina	brain (corpora quadrigemina)	-	adult	male	normal
AA560524	Stratagene mouse Tcell 937311	lymphatic system	T-cells (M30 CD4+ cells)	-	-	normal
AW320348	NCI_CGAP_Mam6	mammary gland	-	5 months	-	infiltrating ductal carcinoma
BB651685	RIKEN full-length enriched, ES cells	-	embryonic stem cells	embryo	-	normal
BB648263	RIKEN full-length enriched, 16 days embryo head	head	-	embryo, 16 day post conception	-	normal
BB624291	RIKEN full-length enriched, 15 days embryo male testis	testis	-	embryo, 15 day post conception	male	normal
BE373165	NCI_CGAP_Mam1	mammary gland	biopsy sample	3 months, virgin	female	tumor
BB641136	RIKEN full-length enriched, 10 days neonate cortex	brain (cortex)	-	neonate, 10 day	-	normal
L1Tf Antisense Transcripts Ending in Promoter						
AI508188	Soares mouse placenta 4NbMP13.5 14.5	placenta	-	adult	-	-
L1Tf Unknown Direction						
BI988934	Mouse E14.5 retina lambda ZAP II Library	eye	neural retina	embryonic day 14.5 post-fertilization	-	normal
D21794	mouse embryonal carcinoma cell line F9	embryo	embryonal carcinoma F9 cell	embryo	-	embryonal carcinoma
B1 Sense Transcripts Starting in Promoter						
BB666320	RIKEN full-length enriched, 2 days pregnant adult female ovary	ovary	-	adult, 2 days pregnant	female	normal
BE332483	Soares_NMEBA_branchial_arch	embryo	branchial arches	embryo, 10.5 day post conception	-	normal
BI319923	Kaestner ngn3 wt	pancreas	-	embryo, 14.5 day	-	normal
AA915643	Soares_thymus_2NbMT	thymus	-	4 weeks	male	normal
AA475851	Soares_mammary_gland_NbMMG	mammary gland	-	4 weeks	male	normal
AA146223	Soares_thymus_2NbMT	thymus	-	4 weeks	male	normal

BB631271	RIKEN full-length enriched mouse cDNA library, C57BL/6J thymus 16 days neonate	thymus	-	16 days neonate	-	-
BG076607	NIA Mouse 15K cDNA Clone Set	embryo (pooled)	-	pooled	-	normal
BF468436	NIH_BMAP_Ret2	eye (retina)	-	neonate, 1 day	-	normal
AI614308	Stratagene mouse skin (#937313)	skin (whole)	-	11 weeks	female	normal
B1 Sense Transcripts Ending in Promoter						
BB733108	RIKEN full-length enriched, 12 days embryo whole body	embryo (whole)	-	embryo (12 day post conception)	-	normal
BB727704	RIKEN full-length enriched, 8 cells embryo	-	-	embryo (8 cell)	-	-
BB696107	RIKEN full-length enriched, 2 days neonate sympathetic ganglion	nervous tissue	sympathetic ganglion	neonate, 2 day	-	normal
BB735303	RIKEN full-length enriched, 6 days neonate spleen	spleen	-	neonate, 6 day	-	normal
BB807678	RIKEN full-length enriched, 13 days embryo brain	brain	-	13 days embryo	-	-
BB697799	RIKEN full-length enriched, 2 days neonate sympathetic ganglion	nervous tissue	sympathetic ganglion	neonate, 2 day	-	normal
BB691116	RIKEN full-length enriched, 2 days neonate sympathetic ganglion	nervous tissue	sympathetic ganglion	neonate, 2 day	-	normal
BB690776	RIKEN full-length enriched, 12 days embryo female mullerian duct	mullerian duct	-	embryo, 12 day post conception	female	normal
BB690518	RIKEN full-length enriched, 12 days embryo female mullerian duct	mullerian duct	-	embryo, 12 day post conception	female	normal
BB683017	RIKEN full-length enriched, 12 days embryo female mullerian duct	mullerian duct	-	embryo, 12 day post conception	female	normal
B1 Antisense Transcripts Starting in Promoter						
BM224073	NIA Mouse Hematopoietic Stem Cell (Lin-/c-Kit+/Sca-1-) cDNA Library (Long)	lymphoreticular tissue	hematopoietic stem cell (Lin-/c-Kit+/Sca-1-)	Age approx.10 weeks old	-	normal
AW541322	NIA Mouse E7.5 Extraembryonic Portion cDNA Library	extraembryonic portion	-	embryo, 7.5 day post conception	-	normal
BG311943	Melton Mouse Adult Pancreas 1	pancreas (total)	-	adult	male	normal

BB746348	RIKEN full-length enriched, adult male kidney	kidney	-	adult	male	normal
BB693800	RIKEN full-length enriched, 2 days neonate sympathetic ganglion	nervous tissue	sympathetic ganglion	neonate, 2 day	-	normal
B1 Antisense Transcripts Ending in Promoter						
AA575771	Knowles Solter mouse blastocyst B1	embryo	blastocyst	embryo, post-implantation	-	normal
BB842555	RIKEN full-length enriched, 6 days neonate spleen	spleen	-	neonate, 6 day	-	normal
BB634890	RIKEN full-length enriched, 0 day neonate thymus	thymus	-	neonate, 0 day	-	normal
AA162296	Stratagene mouse skin (#937313)	skin (whole)	-	11 weeks	female	normal
AA815883	Knowles Solter mouse blastocyst B3	embryo	blastocyst	embryo, post-implantation	-	normal
W14813	Soares mouse p3NMF19.5	-	-	fetus, 19.5 day post conception	-	normal
BB634876	RIKEN full-length enriched, 0 day neonate thymus	thymus	-	neonate, 0 day	-	normal
BE289982	NCI_CGAP_Mam5	mammary gland	gross tissue	7 months	female	tumor
AA241211	GuayWoodford Beier mouse kidney day 0	kidney	-	neonate, 0 day	-	normal
AA606710	Knowles Solter mouse blastocyst B1	embryo	blastocyst	embryo, post-implantation	-	normal
B1 Unknown Direction						
AV009092	Mus musculus 18-day embryo C57BL/6J	-	-	embryo, 18 day post conception	-	normal
AV025745	Mus musculus adult C57BL/6J lung	lung	-	adult	-	normal
C89283	Mouse early blastocyst cDNA	-	-	blastocyst	-	normal
AV005815	Mus musculus C57BL/6J heart	heart	-	-	-	normal
AV026744	Mus musculus adult C57BL/6J liver	liver	-	adult	-	normal
AV008121	Mus musculus 18-day embryo C57BL/6J	-	-	embryo, 18 day post conception	-	normal
AV036376	Mus musculus adult C57BL/6J placenta	placenta	-	adult	female	normal

AV005645	Mus musculus C57BL/6J heart	heart	-	-	-	normal
BG807475	Mouse E14.5 retina lambda ZAP II Library	eye	neural retina	embryonic day 14.5 post-fertilization	-	normal
AV286600	RIKEN full-length enriched, 11 days pregnant adult female ovary and uterus	ovary and uterus	-	adult, 11 days pregnant	female	normal
B2 Sense Transcripts Starting in Promoter						
BB625522	RIKEN full-length enriched mouse cDNA library, C57BL/6J epididymis male adult	epididymis	-	adult	male	-
BB643272	RIKEN full-length enriched, 9.5 days embryo parthenogenote	-	parthenogenote	embryo, 9.5 day post conception	-	normal
BB637459	RIKEN full-length enriched, adult male aorta and vein	aorta and vein	-	adult	male	normal
BB613244	RIKEN full-length enriched, 10 day neonate skin	skin	-	neonate, 10 day	-	normal
BB662937	RIKEN full-length enriched, 15 days embryo head	head	-	embryo, 15 day post conception	-	normal
BB664733	RIKEN full-length enriched, 0 day neonate eyeball	eye	-	neonate, 0 day	-	normal
BB632289	RIKEN full-length enriched, adult male hypothalamus	brain (hypothalamus)	-	adult	male	normal
AI508468	Soares mouse lymph node NbMLN	lymph node	-	4 weeks	male	normal
AA874146	Stratagene mouse skin (#937313)	skin (whole)	-	11 weeks	female	normal
AA727577	Stratagene mouse skin (#937313)	skin (whole)	-	11 weeks	female	normal

B2 Sense Transcripts Ending in Promoter						
BB692532	RIKEN full-length enriched, 2 days neonate sympathetic ganglion	nervous tissue	sympathetic ganglion	neonate, 2 day	-	normal
BB682457	RIKEN full-length enriched, 12 days embryo male wolffian duct	-	wolffian duct includes surrounding region	embryo, 12 day post conception	male	normal
BB007279	RIKEN full-length enriched, 10 day neonate skin	skin	-	neonate, 10 day	-	normal
BB307469	RIKEN full-length enriched, adult male corpora quadrigemina	brain (corpora quadrigemina)	-	adult	male	normal
BB232978	RIKEN full-length enriched, 3 days neonate thymus	thymus	-	neonate, 3 day	-	normal
BB437436	RIKEN full-length enriched, 9 days embryo	embryo	-	embryo, 9 day post conception	-	normal
BB307806	RIKEN full-length enriched, adult male corpora quadrigemina	brain (corpora quadrigemina)	-	adult	male	normal
BB137894	RIKEN full-length enriched, adult male bone	bone	-	adult	male	normal
BB730416	RIKEN full-length enriched, 8 cells embryo	-	-	embryo (8 cell)	-	-
BB725901	RIKEN full-length enriched, 8 cells embryo	-	-	embryo (8 cell)	-	-
B2 Antisense Transcripts Starting in Promoter						
BI076781	NIA Mouse Newborn Ovary cDNA Library	ovary	-	newborn	female	normal
BE990618	NIH_BMAP_MHI2_S1	brain (hippocampus)	-	neonate, 27-32 days	-	normal
BM235719	NIA Mouse Mesenchymal Stem Cell cDNA Library (Long)	-	mesenchymal stem cells	-	-	-
AI315173	Sugano mouse kidney mkia	kidney	-	adult	female	normal
AW113423	mouse liver, vehicle control	liver	-	-	male	normal
BM233927	NIA Mouse Osteoblast cDNA Library (Long)	bone	osteoblasts (KUSA-A1 cells)	-	-	normal
BM228180	NIA Mouse Unfertilized Egg cDNA Library (Long)	egg (unfertilized)	-	egg (unfertilized)	-	normal

BM225535	NIA Mouse Unfertilized Egg cDNA Library (Long)	egg (unfertilized)	-	egg (unfertilized)	-	normal
BB165558	RIKEN full-length enriched, 16 days neonate thymus	thymus	-	neonate, 16 day	-	normal
AU023235	Mouse unfertilized egg cDNA	egg	-	egg, unfertilized	-	normal
B2 Antisense Transcripts Ending in Promoter						
AA733543	Stratagene mouse skin (#937313)	skin (whole)	-	11 weeks	female	normal
AA137521	Barstead MPLRB1	kidney	-	6 weeks	-	normal
AA592590	Barstead mouse irradiated colon MPLRB7	colon	-	embryo, 8 weeks	-	-
BQ044604	NIH_BMAP_EF0	brain (whole)	-	embryo 18.5 dpc	-	normal
AA270323	Soares mouse 3NME12 5	embryo	-	embryo, 12.5 day post conception total fetus	-	-
BB839783	RIKEN full-length enriched, 12 days embryo whole body	embryo (whole)	-	embryo (12 day post conception)	-	normal
BG076961	NIA Mouse 15K cDNA Clone Set	embryo (pooled)	-	pooled	-	normal
AU079450	Sugano mouse brain mncb	brain	-	adult	female	normal
AA624328	Knowles Solter mouse blastocyst B1	embryo	blastocyst	embryo, post-implantation	-	normal
AA671432	Knowles Solter mouse blastocyst B1	embryo	blastocyst	embryo, post-implantation	-	normal
B2 Unknown Direction						
AV004059	Mus musculus C57BL/6J kidney	kidney	-	-	-	normal
C88546	Mouse early blastocyst cDNA	-	-	blastocyst	-	normal
AV121199	Mus musculus C57BL/6J 10-day embryo	-	-	embryo, 10 day post conception	-	normal
BG803328	Mouse E14.5 retina lambda ZAP II Library	eye	neural retina	embryonic day 14.5 post-fertilization	-	normal
BG805335	Mouse E14.5 retina lambda ZAP II Library	eye	neural retina	embryonic day 14.5 post-fertilization	-	normal
BG801033	Mouse E14.5 retina lambda ZAP II Library	eye	neural retina	embryonic day 14.5 post-fertilization	-	normal

BI990347	Mouse E14.5 retina lambda ZAP II Library	eye	neural retina	embryonic day 14.5 post-fertilization	-	normal
BE192844	R3TA	adipose tissue	undifferentiated 3T3 cell line	-	-	-
BE192841	R3TA	adipose tissue	undifferentiated 3T3 cell line	-	-	-
AV077518	Mus musculus stomach C57BL/6J adult	stomach	-	adult	male	normal
B3 Sense Transcripts Starting in Promoter						
BB654601	RIKEN full-length enriched, 9 days embryo	embryo	-	embryo, 9 day post conception	-	normal
BF023128	Soares_thymus_2NbMT	thymus	-	4 weeks	male	normal
AA207499	GuayWoodford Beier mouse kidney day 7	kidney	-	juvenile, 7 day old	-	normal
BB852496	RIKEN full-length enriched, B16 F10Y cells	-	B16 F10Y cells	-	-	-
AI463736	Soares mouse lymph node NbMLN	lymph node	-	4 weeks	male	normal
W11199	Soares mouse p3NMF19.5	-	-	fetus, 19.5 day post conception	-	normal
AA415611	Knowles Solter mouse 2 cell	embryo	-	embryo, 2-cell stage	-	normal
BF016865	Soares_thymus_2NbMT	thymus	-	4 weeks	male	normal
AI585636	Stratagene mouse melanoma (#937312)	skin	M2 cells	-	-	melanoma
AI157954	Soares_thymus_2NbMT	thymus	-	4 weeks	male	normal
B3 Sense Transcripts Ending in Promoter						
BB734748	RIKEN full-length enriched, 6 days neonate spleen	spleen	-	neonate, 6 day	-	normal
BM249345	NIA Mouse 8.5-dpc Whole Embryo cDNA Library (Long)	embryo (whole) including extraembryonic tissu	-	8.5-days postcoitum	-	normal
BB837112	RIKEN full-length enriched, mammary gland RCB-0527 Jyg-MC(B) cDNA	-	mammary gland	-	-	-

BB828833	RIKEN full-length enriched, mammary gland RCB-0527 Jyg-MC(B) cDNA	-	mammary gland	-	-	-
AV252993	RIKEN full-length enriched, 0 day neonate head	head	-	neonate, 0 day	-	-
AV228934	RIKEN full-length enriched, 0 day neonate skin	skin	-	neonate, 0 day	-	normal
BB719109	RIKEN full-length enriched, adult male liver tumor	liver	-	adult	male	tumor
BB066848	RIKEN full-length enriched, 15 days embryo male testis	testis	-	embryo, 15 day post conception	male	normal
BB551130	RIKEN full-length enriched, 2 days pregnant adult female oviduct	oviduct	-	adult, 2 days pregnant	female	normal
AI596376	Knowles Solter mouse blastocyst B1	embryo	blastocyst	embryo, post-implantation	-	normal
B3 Antisense Transcripts Starting in Promoter						
BG074583	NIA Mouse 15K cDNA Clone Set	embryo (pooled)	-	pooled	-	normal
B3 Antisense Transcripts Ending in Promoter						
AA511827	Beddington mouse embryonic region	embryo	-	embryo, 7.5 day post conception	-	normal
BE655207	NIH_BMAP_MSC_N	spinal cord	-	27-32 days	-	normal
BG081998	NIA Mouse 15K cDNA Clone Set	embryo (pooled)	-	pooled	-	normal
BG084130	NIA Mouse 15K cDNA Clone Set	embryo (pooled)	-	pooled	-	normal
BG088273	NIA Mouse 15K cDNA Clone Set	embryo (pooled)	-	pooled	-	normal
BG082673	NIA Mouse 15K cDNA Clone Set	embryo (pooled)	-	pooled	-	normal
BG867898	NCI_CGAP_SG2	salivary gland	-	-	-	normal
BF729691	Soares_thymus_2NbMT	thymus	-	4 weeks	male	normal
BE634061	Soares mouse 3NbMS	spleen	-	4 weeks	male	normal
BB647536	RIKEN full-length enriched, 10 days neonate cerebellum	brain (cerebellum)	-	neonate, 10 day	-	normal

B3 Unknown Direction						
AV150616	Mus musculus hippocampus C57BL/6J adult	brain (hippocampus)	-	adult	male	normal
AV135910	Mus musculus C57BL/6J 10-11 day embryo	-	-	embryo, 10-11 day post conception	-	normal
AV082796	Mus musculus tongue C57BL/6J adult	tongue	-	adult	male	normal
AV077428	Mus musculus stomach C57BL/6J adult	stomach	-	adult	male	normal
BG807953	Mouse E14.5 retina lambda ZAP II Library	eye	neural retina	embryonic day 14.5 post-fertilization	-	normal