NOVEL GENETIC EFFECTS OF A HUMAN ENDOGENOUS RETROVIRUS INSERTION

by

PAUL EDWARD KOWALSKI

B.Sc. (Honours Biology), University of Waterloo, 1992

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

In

THE FACULTY OF GRADUATE STUDIES

Department of Medical Genetics

We accept this thesis as conforming to the required standard

THE UNIVERSITY OF BRITISH COLUMBIA

September, 1998

© Paul Edward Kowalski, 1998

In presenting this thesis in partial fulfilment of the requirements for an advanced degree at the University of British Columbia, I agree that the Library shall make it freely available for reference and study. I further agree that permission for extensive copying of this thesis for scholarly purposes may be granted by the head of my department or by his or her representatives. It is understood that copying or publication of this thesis for financial gain shall not be allowed without my written permission.

Department of MEDICAL GENETICS

The University of British Columbia
Vancouver, Canada

Date 21/SEPT/98

DE-6 (2/88)

Abstract

Human endogenous retroviruses (HERVs) are repetitive, noninfectious chromosomal elements degenerated from exogenous retroviruses, and compose as much as 2% of the human genome. The HERV-H family numbers approximately 1000 elements dispersed throughout the human genome. HERV-H elements have been shown to affect the expression of adjacent cellular genes. For example, in teratocarcinoma cell lines, a HERV-H LTR promotes expression of, and splices into a downstream cellular transcript, PLA2L, which contains two phospholipase A_2 (PLA_2)-like domains. PLA2L was determined to be a tripartite fusion transcript, composed of HERV-H sequences, 8-10 exons of an unknown but conserved gene HHAG-1(HERV-H associated gene 1), and a downstream gene encoding an inner ear structural protein, termed otoconin-90. As no chromosomal rearrangements were found in the teratocarcinoma cell lines expressing the PLA2L fusion, intergenic splicing influenced by the HERV-H promoter is hypothesized to be the cause of gene fusion. Cloning and characterization of both the human genomic locus and the murine otoconin-90 cDNA confirmed that PLA2L is a fusion transcript. The HERV-H insertion into an intron of the HHAG-1 gene was determined to have occurred 15-20 million years ago, with the HERV-H element in this locus being stable and present in all humans and higher primates. The region was localized to human chromosome 8q24.1-8q24.3. Although the tripartite transcript is abundant in teratocarcinoma cell lines, no evidence of protein synthesis was detected in teratocarcinoma cell lysates. Heterologous expression experiments have shown that the full-length HERV-H-containing cDNA is transcribed but not translated in COS cells. However, a 5' deletion construct which removes the HERV-H-encoded sequence is efficiently translated,

while both constructs were transcribed at comparable levels. These effects are postulated to be caused by the HERV-H sequences acting as a translational inhibitory type of 5' UTR, containing elements known to repress protein synthesis. Both the translation-level effect of a HERV upon an adjacent gene and a HERV-H-associated intergenic fusion have not been previously reported, and suggest more complex types of effects which HERV elements can exert upon nearby human genes.

Table of Contents

Abstract ii

List of Tables vii

List of Figures viii

List of Abbreviations x

Acknowledgements xi

Chapter One : Introduction 1

1.1 Repetitive Elements in the Human Genome 2

1.2 Short Interspersed Nuclear Repeats (SINEs) 3

1.3 Long Interspersed Nuclear Elements (LINEs) 4

1.4 Human Endogenous Retroviruses (HERVs) 5

1.4.1 Relationship to exogenous retroviruses 5

1.4.2 Endogenous Retroviruses (ERVs) 8

1.4.3 Potential genetic effects of retroviral insertions 12

1.4.4 Cellular effects of ERV expression 16

1.5 HERV effects upon adjacent cellular genes and disease 21

1.6 HERV-H and effects upon heterologous genes 23

1.7 Rationale and Thesis Objectives 31

Chapter Two : Materials and Methods 33

2.1.1 Library Screening and Genomic Cloning 34

2.1.2 Evolutionary Genomic PCR and Chromosomal Mapping 36

2.1.3 Cell Lines and Nucleic Acid /Protein Extractions 37

2.1.4 Southern and northern blotting and hybridizations 38

2.1.5 Murine cDNA synthesis and RT-PCR 40

2.1.6 Construction of full length murine otoconin-90 cDNA 41

2.1.7 Plasmid DNA isolation and sequencing 42

2.1.8 Expression of PLA2L:GST fusion proteins and generation of anti-PLA2L antiserum 43

2.1.9 Western blotting and probing 45

2.1.10 Transfections 45

2.1.11 Construction of PLA2L Expression Vectors 46

2.1.12 Anti-Thy-1 Flow Cytometry 48

Chapter Three : Genomic Structure and Evolution of the Human PLA2L Locus 49

3.1 Introduction 50

3.2 Results 54

3.2.1 PLA2L Genomic Cloning 54

3.2.2 Intron-Exon Structure 56

3.2.3 Age of the HERV-H insertion 62

3.2.4 Chromosomal Localization of PLA2L 66

3.2.5 Evolutionary conservation of PLA2L 68

3.3 Discussion 70

Chapter Four: cloning and characterization of the murine homologue of human otoconin-90, and development of a hypothesis for PLA2L biogenesis 75

4.1 Introduction 76

4.2 Results 80

4.2.1 Identification and genomic cloning of murine homologue 80

4.2.2 Construction of an otoconin-90 murine genomic contig 81

4.2.3 cDNA cloning and consensus sequence assembly 81

4.2.4 Murine otoconin-90 has an independent and divergent 5' end 92

4.2.5 Prediction of secretion signals in analogous regions of murine otoconin-90 and human PLA2L 93

4.2.6 Cloning of human intergenic genomic region 99

4.3 Discussion 103

4.3.1 Otoconin-90 expression 103

4.3.2 Structural implications of conserved PLA-domains 105

4.3.3 Antitermination and intergenic splicing 108

Chapter Five: HERV-H suppresses translation of an associated fusion transcript, PLA2L 113

5.1 Introduction 114

5.2 Results 115

5.2.1 Expression and purification of PLA2L fusion proteins 115

5.2.2 Endogenous expression of PLA2L in teratocarcinoma cells 117

5.2.3 HERV-H sequences affect translation of the PLA2L mRNA. 120

5.2.4 HERV-H sequences suppress PLA2L translation, not transcription. 122

5.2.5 HERV-H sequences do not suppress translation of a heterologous gene, Thy-1 123

5.3 Discussion 128

Chapter Six: Summary and conclusion 133

References 141

List of Tables

Table 1 Some important HERV families 11

Table 2 HERV effects upon adjacent human genes 22

Table 3 PCR Primers Used 37

Table 4. Exon-Intron Boundaries and Exon and Intron Sizes of PLA-domains of PLA2L 59

List of Figures

Figure 1.1 Schematic structure of an integrated retrovirus 6

Figure 1.2 Potential effects of ERVs upon adjacent gene expression 13

Figure 3.1 Schematic structures of the PLA2L cDNA and probes. 53

Figure 3.2 Genomic map of the human PLA2L locus. 57

Figure 3.3 Similarity of PLA-domains to $sPLA_2$ and intron/exon structure. 60

Figure 3.4 Integration time of HERV-H into the PLA2L locus. 63

Figure 3.5 Primate speciation and radiation relative to HERV-H expansion. 65

Figure 3.6 Chromosomal and regional localization of PLA2L. 67

Figure 3.7 DNA sequence conservation of PLA2L. 69

Figure 4.1 Schematic of murine otoconin-90 discovery and relation to PLA2L. 79

Figure 4.2 Assembly of murine otoconin-90 cDNA contig. 83

Figure 4.3 Otoconin-90 cDNA sequence. 84

Figure 4.4 Alignment of murine and human otoconin-90 DNA sequence 88

Figure 4.5 Amino acid alignment of murine otoconin-90 and human PLA2L 89

Figure 4.6 Alignment of otoconin PLA-domains 90

Figure 4.7 Predicted secretion signals in otoconin-90 and PLA2L 97

Figure 4.8 Localization of signal peptides in PLA2L and otoconin-90 proteins 98

Figure 4.9 Original and revised composite PLA2L structure 101

Figure 4.10 Detailed schematic map of human intergenic region 102

Figure 5.1 Schematic and sequence of the 5' region of the PLA2L fusion transcript. 116

Figure 5.2 Expression and purification of PLA2L fusion proteins. 119

Figure 5.3 Anti-PLA2L western blot of teratocarcinoma lines and PLA2L transfectants. 121

Figure 5.4 Northern blot of PLA2L transfectants. 124

Figure 5.5 FACS dotplots of HERV-H/Thy-1 chimera transfectants. 127

Figure 5.6 Translational inhibitory structures within the PLA2L 5'UTR 130

Figure 6.1 How HERV-H affects the PLA2L locus 135

List of Abbreviations

CMV	cytomegalovirus
DMEM	Dulbecco's modified Eagle media
dNTP	deoxynucleotide
E14	embryonic day 14 (murine)
EDTA	(ethylenedinitrilo)tetraacetic acid
EST	expressed sequence tag
GCG	Genetics Computer Group
GSH	glutathione
GST	glutathione-S-transferase
HDTV	human teratocarcinoma-derived virus
HERV	human endogenous retrovirus
HHAG-1	HERV-H-associated gene 1
HIV	human immunodeficiency virus
HMG	high mobility group
IAP	intracisternal A particle
IDDM	insulin dependent diabetes mellitus
kb	kilobases
kDa	kilodaltons
LINE	long interspersed nuclear element
Mb	megabases
μORF	micro open reading frame
MS	multiple sclerosis
MSRV	multiple sclerosis-associated retrovirus
MYA	millions of years ago
ORF	open reading frame
PAGE	polyacrylamide gel electrophoresis
PLA	phospholipase A_2
PLA2L	PLA_2-like gene
PLT	placental LTR terminated gene
RACE	rapid amplification of cDNA ends
RNAPII	RNA polymerase II
RT	reverse transcriptase
SAG	superantigen
SD	splice donor
SDS	sodium dodecyl sulfate
SINE	small interspersed nuclear element
snRNP	small nuclear ribonucleoprotein
spp.	specific
$sPLA_2$	secreted phospholipase A_2
SS	signal sequence
TAE	Tris-acetate-EDTA
TBE	Tris-borate-EDTA
TC	teratocarcinoma
UTR	untranslated region
UV	ultraviolet

Acknowledgements

First and foremost I'd like to gratefully thank my Ph.D. supervisor Dixie Mager. Her unfailing encouragement, mentorship and support contributed more than anything else to my doctoral research. I would be a far lesser scientist without your guidance.

If not for my parents enduring love, support, and specifically their drive to always provide the best opportunities for their children, I would never have made it to graduate school. Thanks, Mom and Dad. This thesis is dedicated to you.

Without the love and encouragement of my Aunt Kris, this Ph.D. might have been in something far removed from Medical Genetics- like history! You made me interested in science, at an early age.

The Mager Lab members, past and present, have made my time there very enjoyable, almost without exception. David Wilkinson, Nancy Cooper, Jacques Brennan and Soheyl "Soldier" Baban taught me a lot, and were great lab-mates. Dave Nelson didn't teach me much about science, but a great deal about everything else, such as mountain biking, headlight design, beer drinking, explosives etc. Dave's is one of the excellent friends I've made in the Mager lab, the other being Karina McQueen. Winner of "Miss Personality" awards 1995-98 inclusive, Karina puts up with me soaking her with water, stealing her solutions etc., and still maintains her sunny disposition! Thanks Karina, the last 3 years would have been pretty dull without you. Thanks also to Patrik Lutefisk Medstrand, for his helpfulness, friendship and voluminous HERV knowledge. Last but definitely not least, many thanks are due to The Supreme Technician Doug Freeman. Doug has patiently taught me a great deal about the day-to-day methods of molecular biology. Without your help, my thesis would have taken even longer to finish!

I made many friends at the TFL, foremost and best among them Laurie Ailles. Her friendship, caring and patience both inside and outside the lab (and all over the west of Canada and the US) have made the "Vancouver Years" much more enjoyable. Thanks for everything- I'll miss you!

Adding an much-needed bit of outright freakishness, Mark "Shaft" Ware has also been a great guy and a good friend, and my protein work and snowboarding wouldn't have been the same (or possible, more likely) without him. In addition to being responsible for my addiction to coffee, the Usual Suspects at "the Garden" made life (and mornings) much more fun over the past few years. Thanks Ron, Laurie, Ying (The Doctor Coffee), and Doug. You're all great friends! Space doesn't permit thanking all the friends at TFL and UBC, but I want to specially thank Heather, MTL, Cristina, Patty, Carmine, Tracy D and Tracy S, Christine, Sarah, Rob, Sharon, Saghi, Lori, Daryl, Grant, Mana, Steph and Philip.

Finally, my Thesis Committee (Ann Rose, Ross MacGillivray and Rob Kay) deserve thanks for steering me in the right direction through the years. Special credit is due to Ann for her many thoughtful suggestions for my final thesis, and to Rob for being a "second mentor" and for his continued excellent help with experimental design.

CHAPTER ONE : INTRODUCTION

1.1 Repetitive Elements in the Human Genome

Mammalian genomes are characteristically gene-poor, and dominated by repetitive, non-coding sequences separating widely-spaced transcriptional units. This is in contrast to the compact, gene-dense prokaryotic genomes which are predominantly composed of protein-coding regions (Henikoff et al. 1997). These repetitive sequences can be placed into two general categories: simple sequence repeats such as α-satellite centromeric DNA, and interspersed genome-wide repetitive DNA derived from transposable elements. The latter class has been implicated in playing a major role in the evolution of the mammalian genome, refashioning the genomic architecture via the facilitation of translocations, gene duplications and conversions, and heterologous recombination (Lindahl 1991; O'Neill et al. 1998). Additionally, some elements may have evolved to comprise tissue-specific gene promoters or enhancers (Britten 1996).

The fraction of the human genome composed of transposon-derived repeats has recently been estimated to be 35%, as extrapolated from 7 Mb of contiguous human genomic sequence (Smit 1996; Henikoff et al. 1997). The actual fraction could be even higher, taking into account highly degenerate repeat sequences. Human transposon-derived repeat elements are subdivided into four categories: short interspersed nuclear elements (SINEs), long interspersed nuclear elements (LINEs), and endogenous retrovirus elements (ERVs), all of which are classified as retroelements, or elements which are derived from reverse transcription of RNA. The fourth category of transposon-derived repeats are remnants of DNA transposons, which encoded (or contain sequences derived from) transposases, moved by DNA excision, and are related to the *Ac/hobo* and *Tc1/mariner* classes of yeast/insect

transposons (Smit 1996). Less than 2% of the above fraction is derived from DNA transposons, with the bulk of the interspersed repeat fraction of the human genome composed of reverse-transcribed RNA. This bias is not surprising considering that reverse transcription of retroelement RNA and subsequent reintegration is an inherently duplicative and additive process.

1.2 Short Interspersed Nuclear Repeats (SINEs)

Surprisingly, the fraction of the genome devoted to RNA-derived sequences far exceeds that of protein-encoding exons (Smit 1996). This may be due to the enigmatic role they play in genome structure and evolution, and likely is also a result of the inability of the host genome to remove these intragenomic "parasites", leading to their continuous increase throughout evolution. By far the most numerous, and therefore replicatively successful of the retroelements are SINEs, which include Alu elements, and MIR elements in human and B1/B2 repeats in mice. In human, they number an estimated 1.6 million copies in total. These elements are partly derived from tRNAs or 7SL RNAs (a component of the signal recognition particle in the endoplasmic reticulum), from which they obtain their internal RNA polymerase III promoter (Labuda et al. 1995). This internal promoter is active in all cell types and can direct position-independent expression, which is proposed to be the reason for the very high copy number of SINEs. They are obliged to use a cellular source of reverse transcriptase as they are not protein coding. The SINE-independent source of reverse transcriptase may be of HERV origin or possibly a telomerase (Eickbush 1997), but recent evidence suggests it arises from LINE elements (Ohshima et al. 1996). Given the very high copy number, it is not surprising that Alu insertions or Alu-mediated recombinations have been shown to be causal in a number of human genetic disorders. Retrotranspositions

into the butyryl cholinesterase and Factor IX genes causing acholinesterasemia and hemophilia B, respectively, are two examples (Muratani et al. 1991; Vidaud et al. 1993). Interestingly, all Alu insertions appear to involve evolutionarily young Alu subfamilies (Labuda et al. 1995). Even more common appear to be Alu-Alu recombinations resulting in genomic deletions, which cause genetic disorders in a number of cases, with a well known example being an Alu-mediated deletion in the low density lipoprotein receptor causing familial hypercholesterolemia (Lehrman et al. 1985).

1.3 Long Interspersed Nuclear Elements (LINEs)

LINE elements are approximately half as numerous as SINEs, with an estimated 870,000 copies present in a 3 billion bp human genome (Smit 1996) but as they are much larger than SINEs, with a unit length of 6-8 kb, they compose a greater proportion of the human genome by mass (16.7% LINEs compared to 11.7% SINEs) (Henikoff et al. 1997). LINE elements are the most active known human retrotransposons (Moran et al. 1996; Sassaman et al. 1997), with *de novo* insertions implicated in various single-gene defect genetic diseases (Miki et al. 1992; Kazazian and Moran 1998). Containing two potential coding regions, LINEs encode a reverse transcriptase (RT)/endonuclease from the second ORF and a partially characterized protein with RNA-binding activity from the first ORF (Feng et al. 1996; Hohjoh and Singer 1997). However, most (95%) LINE elements are truncated at their 5' ends, do not possess open reading frames (ORFs), and are not retrotranspositionally active (Kazazian and Moran 1998). LINE elements, when undeleted, contain an internal RNA polymerase II promoter, ensuring position-independent expression and, like retroviruses, retention of the promoter following retrotransposition. Intact elements

numbering 30-60 per diploid human genome, appear to be "master" elements with uninterrupted ORFs and functional promoters, and presumably act as the source for most new LINE retrotranspositions (Sassaman et al. 1997).

1.4 Human Endogenous Retroviruses (HERVs)

1.4.1 Relationship to exogenous retroviruses

The human genome contains an estimated 50,000 repetitive elements which are related, structurally and by sequence similarity, to the genomes of exogenous or infectious retroviruses (Wilkinson et al. 1994; Lower et al. 1996). Exogenous retroviruses possess a plus-stranded RNA genome which mimics the structure of eukaryotic mRNA, possessing a 5' cap and a 3' poly(A) tail, in order to ensure efficient translation of viral genes by host cell systems. Upon infection and entry into a permissive host cell, this RNA genome is reverse transcribed into double stranded DNA by the action of the retroviral reverse transcriptase enzyme. The DNA genome is then translocated to the nucleus where it is integrated into the host cell genome as a provirus (Coffin 1992). This chromosomal integration tends to occur in transcriptionally active regions of DNA (Patience et al. 1997) and is carried out by the viral integrase enzyme, which possesses a endonucleolytic activity to create the initiating nick in a strand of chromosomal DNA (Coffin 1992). The genome structure of a typical simple retrovirus (C-type) in its integrated, proviral form, is shown in Figure 1.1 (Coffin 1992). The characteristic structure of a retrovirus includes the three main structural

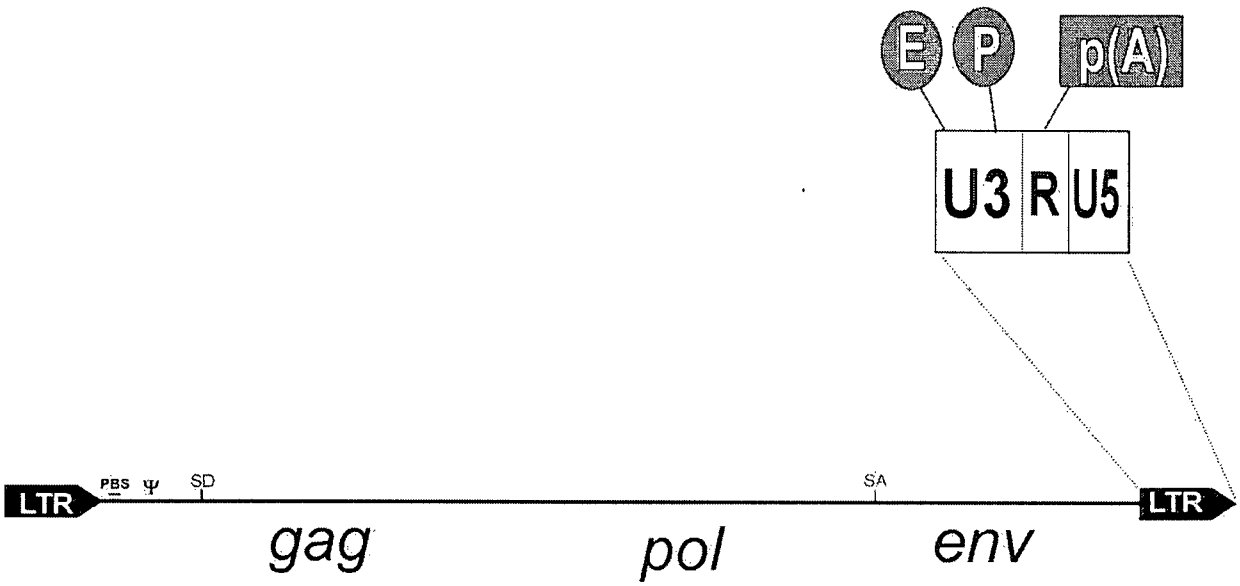


Figure 1.1 Schematic structure of an integrated retrovirus
Prototypical retrovirus proviral structure shown with a magnification of the 3' LTR at top right. Most HERVs are deleted or mutated in the internal regions; for example, most HERV-H elements lack the *env* region and parts of *pol*. Infectious retroviruses possess all major structural genes: *gag, pol,* and *env*. Downstream of the 5' LTR in both HERVs and exogenous retroviruses, the leader region contains the primer binding site (PBS), the Ψ packaging signal, and the major splice donor site (SD). tRNA complementarity in the PBS is used to classify HERVs into families. Retroviruses (but not most HERV-H's) have a splice acceptor (SA) site at the start of the *env* region. LTRs are divided into the U3, R, and U5 regions; R and U5 contain structures required for reverse transcription and are generally much shorter than U3. U3, in both exogenous and endogenous retroviruses contain the enhancer (E) and promoter (P) elements. Polyadenylation signals (p(A)) are generally found within the R region.

transcriptional units, *gag*, *pol* and *env*, flanked by long terminal repeats (LTRs), which are necessary for viral replication and transcription. Retroviral LTRs contain a U3 region, an R region and a U5 region, in 5'→ 3' order. The U3 and U5 regions are unique in the RNA genome but are duplicated in the provirus, whereas the R region is found duplicated in both retroviral genomic RNA and proviral DNA. Composing most of the length of the LTR, the U3 region contains all the elements needed for transcription of proviral DNA, including enhancer sequences, and promoters. The R region possesses repeat elements used in the reverse transcription process, and the poly(A) signal, while U5 provides the sequences needed for initiation of reverse transcription and for integration (Coffin 1992). The 5' LTR is followed immediately by the tRNA primer binding site (PBS), used to prime reverse transcription, and a leader region of variable length. The leader contains the Ψ signal for packaging the genomic RNA into virion particles, and a major splice donor site, used in intragenomic splicing to generate subgenomic mRNAs.

The *gag* gene follows the leader region, and encodes the three structural proteins which form the virion: the matrix, capsid and the RNA-binding nucleocapsid. The small *pro* gene, coding for the protease which cleaves viral polyproteins, is located between the *gag* region and the downstream *pol* region. *Pol* encodes the reverse transcriptase enzyme, the RNase H activity as well as the integrase enzyme. The final primary transcriptional unit is the *env* region, which, coding for the viral envelope glycoproteins, determines the host range and tropism of a retrovirus (Coffin 1992). Additionally, possession of the *env* region distinguishes an exogenous or endogenous retrovirus from a LTR-containing retrotransposon such as the yeast Ty1 element (Garfinkel 1992). While no retrovirus-related LTR-retrotransposon (similar to Ty1/IAP) has been

detected in the human genome, it does possess greater than 150,000 copies of a related but enigmatic family of primate LTR-retrotransposons, the THE-1/MaLR elements (Smit 1996). A THE-1 consensus sequence contains an ORF, but without discernable homology to reverse transcriptase or any retroviral gene.

1.4.2 Endogenous Retroviruses (ERVs)

The prevailing hypothesis explaining the diversity and high copy number of ERVs in the mammalian genome proposes that they arose from ancient germ-line infections by simple exogenous retroviruses, with subsequent retrotransposition and mutation to generate the many non-coding, non-infectious families seen in contemporary genomes (Wilkinson et al. 1994). Integration of retroviruses into germ cell chromosomes, in the absence of lethal insertional mutagenesis, will allow those proviruses to be stably transmitted to offspring as a Mendelian trait. The belief that *all* ERVs are remnants of exogenous retroviruses is challenged by a current hypothesis of retrovirus evolution, which proposes that retroviruses evolved from reverse transcriptase-containing retrotransposons. This implies that some ERV-related elements still present in mammalian genomes might be the progenitors of exogenous retroviruses rather than the opposite (Temin 1992; Lower et al. 1996).

The murine genome contains a much larger number of retrotranspositionally active ERVs than the human, with many of the ERVs appearing to be "endogenized" versions of exogenous viruses, such as mouse mammary tumor virus (MMTV)(Golovkina et al. 1992). When similar endogenous and exogenous viruses are present, the opportunity exists for the endogenous forms to supply viral proteins to infectious forms in *trans*, potentially changing specificities or tropisms of the

exogenous virus, or serving to block cells expressing endogenous retroviral *env* proteins from "superinfection" by the similar exogenous strains (Weiss 1993; Patience et al. 1998). This is not likely to occur in human cells, as HERVs identified to date are similar to simple retroviruses such as mammalian C-type (Class I HERVs, including HERV-H, HERV-E, ERV9, HERV-I and S71) or B- and avian C-type (Class II HERVs, including HERV-K, HML-6 and relatives)(see Table 1). Unlike the murine system, no evidence of HERVs closely related to an infectious human retrovirus has been found (Wilkinson et al. 1994). The approximately 50,000 HERVs in the human genome are currently subdivided into two major classes based on retrovirus homology, and into a total of 16 different families (Medstrand 1996). Details of some characterized HERV families are shown in Table 1. It is notable that most of these 50,000 "elements" are predicted to be solitary LTRs. Effects of murine ERVs upon adjacent genes and the organism as a whole have been extensively studied, with results suggesting that ERVs may play a significant role in the biology of their host. Immunological effects such as protection against infection by similar exogenous retroviruses, immunological tolerance and superantigen-induced autoimmunity have been described (Adachi et al. 1993; Medstrand 1996). Donation of sequences to replication-incompetent exogenous retroviruses resulting in infectious retrovirus with a broadened host range has been shown to occur in several cases, with different ERV/retrovirus combinations (Martinelli and Goff 1990; Golovkina et al. 1997). Insertional activation of oncogenes resulting in tumorigenesis is a well known phenomenon in murine and avian systems, as are expression-level effects of ERV LTRs upon unrelated but adjacent cellular genes (Furter et al. 1989; Nusse 1991; Adachi et al. 1993)

As HERVs are transmitted as an integral part of the human genome, the selective pressure to conserve and maintain intact ORFs encoding proteins essential for extrachromosomal viral replication has been lost. This results in almost all HERVs containing gross deletions or point mutations resulting in stop codons within the *gag-pol-env* regions (Lower et al. 1996). This loss of coding capacity may have been selected for during evolution, as the chance of being fixed within the germ line is much greater when a HERV has lost deleterious or pathogenic function. Following the initial establishment of an element in the germ line, HERV genomes presumably expanded in copy number by retrotransposition although expansion by general genomic duplications or rearrangements has likely also occurred. This retrotransposition requires a HERV-encoded reverse transcriptase/integrase from a rare intact element. Additionally, intra-element recombination causing deletion of internal sequences have generated a large number of solitary HERV LTRs (Wilkinson et al. 1994). It is worthwhile to note that for any given *de novo* germ-line HERV insertion, only beneficial or neutral effects will be conserved in evolution; any deleterious effects such as insertional mutagenesis of an essential gene will cause death of the cell or, rarely, death of the organism. The result of either is that the insertion will not be fixed in the germ line, and will disappear when the individual dies. Somatic HERV retrotranspositions, and their attendant effects, are very likely more frequent but as they are lost upon the death of the individual, their study is nearly impossible.

Table 1 Some important HERV families

HERV	Copy number (solitary LTRs)	Major sites of transcription	Protein produced	Notable effects	Reference
HERV-H	~1000 (~1000)	TC cells, placenta	?	1000's of fusion ESTs	(Mager and Henthorn 1984)
HERV-K	~50 (~2000)	TC cells, placenta, tumor cells	gag, pol, env, cORF	Linked to IDDM, forms HTDV particles	(Tonjes et al. 1996)
HERV-E	35-50	Placenta, tumor lines	?	Tissue-spp enhancer of pleiotrophin, amylase	(Ting et al. 1992)
ERV9	~40 (~4000)	Placenta, tumor lines	?	Related to MS-associated HERV	(Lania et al. 1992)
HERV-R	1 (~10)	Placenta, skin, trophoblast	env	High levels of env protein in trophoblast	(Venables et al. 1995)
HML-6	30-40 (~100)	Lung, placenta	?	Promotes HLA-DRB6	(Medstrand et al. 1997)

1.4.3 Potential genetic effects of retroviral insertions

Retroviral insertions can potentially affect adjacent cellular genes in a number of ways as shown in Figure 1.2. Specific examples are given in a following section. Exogenous retroviruses show a preference for integration in transcriptionally active regions, i.e. near genes (Fan 1994), and while not studied in detail, endogenous retroviruses may recapitulate this behavior (Taruscio and Manuelidis 1991). This is likely due to such regions being the only ones "accessible" to the retroviral integrase activity (Patience et al. 1997). Additional indirect evidence for this is the non-random clustering of different types of HERVs seen in various genomic regions (Wilkinson et al. 1994; Lindeskog et al. 1998), such as the hybrid HERV-E.PTN element, derived from recombination between a HERV-E and a HERV-I element, found within an intron in the 5' end of the human pleiotrophin gene (Schulte and Wellstein 1998), or the cluster of HERV-I elements in the haptoglobin locus (Maeda and Kim 1990). Integration within an exon is very disruptive to an ORF; a premature nonsense or termination codon is the anticipated result. This type of integration is very unlikely due to the paucity of exons relative to intronic, or non-coding, DNA. The usual type of integration event will insert a retroviral element into genomic DNA upstream of, downstream of, or in an intron of a cellular gene. Cryptic splicing of intron-based ERV fragments into existing ORFs often introduces stop codons, as most ERV genomes do not contain ORFs. The manner of effect exerted upon an adjacent gene is dependent upon the site of insertion, and is due to the transcriptional and post-transcriptional signals (and, rarely, translational: see Chapter 5) within ERVs: promoters, enhancers, polyadenylation signals and splice donor or acceptor sites.

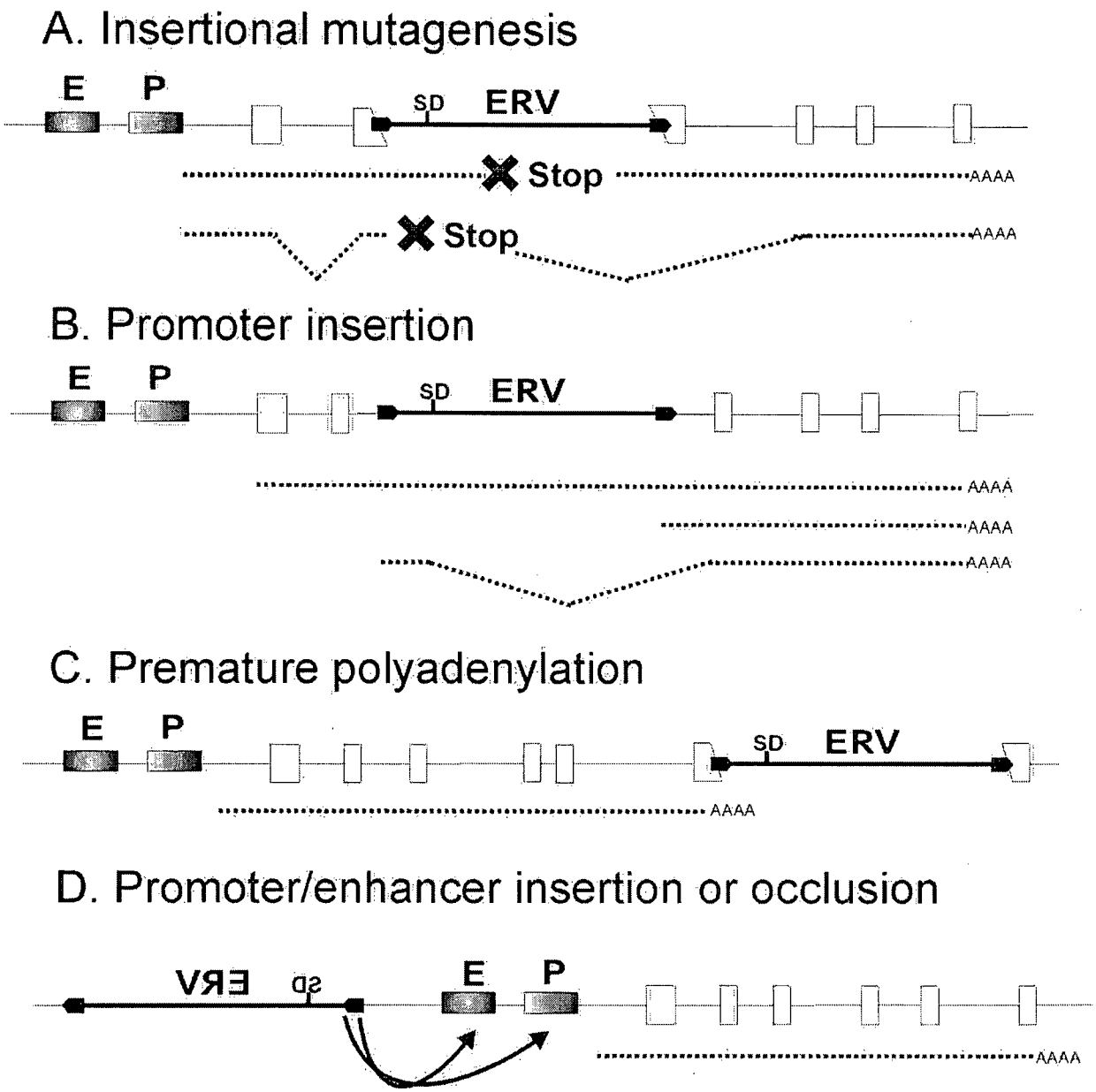


Figure 1.2 Potential effects of ERVs upon adjacent gene expression
The primary types of effects exerted by ERV insertions into unrelated cellular genes are shown above. Panel A shows insertional mutagenesis by introduction of premature stop codon with the ERV integrating within an exon, or being spliced into exon (2nd transcript, panel A). Exons are shown as white boxes, genomic DNA as a thin line, with predicted transcripts shown as dotted lines, below exons. Splicing between exons is not shown. Cellular gene enhancer and promoter shown as grey rectangles with E or P, respectively. ERV element shown in bold, flanked by two LTRs. SD, major internal splice donor site. Translation stop codon shown by a bold X.

Many examples of the effects of ERVs upon adjacent genes are known in the murine system. Apart from the status of the mouse as the primary animal model for experimental genetics, this is likely due to the high retrotranspositional activity of murine ERVs, such as IAP and ETn elements (Wang et al. 1997). For example, in the autoimmune *lpr* strain of mouse an ETn element has integrated into the second intron of the Fas antigen and causes premature truncation of the Fas mRNA via usage of the LTR poly(A) signal. The disrupted Fas antigen transcript leads to loss of apoptosis and subsequent lymphoproliferation causing an autoimmune disease similar to human lupus erythematosus (Adachi et al. 1993). Other notable examples of IAP:gene interactions include being the basis for the dominant yellow agouti coat color mutants (Perry et al. 1994), and apparently causing growth factor independence in a variety of murine hemopoietic cell lines, by inserting and dominantly activating expression of the cytokines IL-3 and GM-CSF (Wang et al. 1997; Pogue-Geile et al. 1998).

A HERV (or ERV) element or solitary LTR integrated upstream of a gene can cause inappropriate temporal or spatial expression of the gene, due to RNA polymerase II promoters and enhancers present in LTRs (Luciw and Leung 1992). Many of the LTR enhancers are position and orientation independent, and may influence genes located a distance downstream, or on the opposite strand (Fig. 1.2). Promoter interference may also occur when a ERV LTR is found within an internal intron of a gene and occludes the normal promoter, generating a mis-expressed, 5' truncated mRNA. This type of event can also occur via splicing, with transcription initiating in a intron-based LTR and then splicing, by use of a ERV or cryptic splice donor site, into a downstream exon of a cellular gene. It is useful to note that, independent of solitary LTRs, *either* LTR in an intact ERV element can exert

transcriptional effects upon adjacent genes. Similar effects can be mediated by ERV LTR enhancers perturbing normal promoter function of a cellular gene. The well characterized HERV-E / amylase gene example of enhancer insertion will be discussed in a following section (Ting et al. 1992). The other primary posttranscriptional control element present in LTRs is the polyadenylation signal (Figure 1.1)(Coffin 1992). This signal can cause protein truncation mutations due to premature polyadenylation of a transcript. The mechanism by which intron-located LTRs contribute poly(A) signals to the associated genes is unclear, but likely involves cryptic splicing of the LTR to normal exons, or more unlikely, usage of LTR poly(A) signals present in the intronic component of pre-mRNA. Interestingly, current models and mechanisms of 3' end formation / polyadenylation have shown to be linked *in vivo* to splicing of the final (3') exon (Berget 1995; Proudfoot 1996). This exon definition model of splicing and polyadenylation incorporates two experimentally proven findings relevant to HERV LTRs causing premature polyadenylation. The first shows that cleavage and polyadenylation is initiated by interactions between the final 3' splice acceptor site and the poly(A) signal (mediated by U1 snRNP) which define the 3' terminal exon. An LTR poly(A) signal lying downstream of an authentic 5' splice donor would not be recognized as such by the splicing / polyadenylation machinery which, in the presence of the 3' splice acceptor and *adjacent* 5' splice donor, would be expected to splice and not polyadenylate the (internal) exon (Berget 1995; Wahle and Kuhn 1997). Interestingly, recent reports have shown that the relative concentrations of the basal polyadenylation factor CstF-64 regulate the usage of multiple alternative polyadenylation sites in various cell types: when CstF-64 is limiting the strongest poly(A) site is chosen, but when the CstF-64 concentration is elevated, the 5'-most site

is used ("first come, first served")(Colgan and Manley 1997). This predicts that cell types containing non-limiting levels of CstF-64 would favor premature polyadenylation by intronic HERV LTR poly(A) signals. The second relevant component of this model states that, in complete proviruses, the strong splice donor site found in the leader region inhibits the 5' LTR poly(A) site via U1 snRNP interaction (Ashe et al. 1997). This was based on elucidation of the mechanism by which retroviruses (HIV-1) exclusively utilize the downstream, 3' LTR poly(A) signal. These results would predict that solitary HERV LTRs, lacking the major splice donor site, would be far more efficient at inducing premature polyadenylation of proximal genes than full-length elements, and that usage of LTR poly(A) signals in introns would be unlikely when additional splice acceptor-containing exons are present downstream. Premature polyadenylation would be predicted to occur efficiently if a HERV LTR integrated into a 3' UTR, which is often the largest exon, and lacks a downstream splice site. Many examples of polyadenylation of cellular transcripts by HERV elements are known (Wilkinson et al. 1994), but these studies generally examine cDNAs, without investigation of the underlying position of the LTR (or complete HERV) in genomic DNA relative to the cellular gene, nor the precise manner of gene:LTR fusion and subsequent polyadenylation.

1.4.4 Cellular effects of ERV expression

Independent of transcriptional effects upon nearby genes, a substantial body of work implicates HERV-encoded proteins in a variety of cellular and organismal effects. One of the best characterized examples of this is the presence of human teratocarcinoma-derived virus particles (HTDV) in normal placental syncytiotrophoblasts, germ cell tumors, T47D mammary carcinomas and various

teratocarcinoma cell lines (Lower et al. 1996; Patience et al. 1996). These non-infectious, retrovirus-like particles can be visualized with electron microscopy and usually appear to be arrested in the budding stage of viral development. A variety of methods including immunogold cytochemistry and gradient purification of particle-associated RNA have shown that HTDV particles are encoded by HERV-K elements. HERV-K genomes possess the most intact ORFs of any HERV family examined, including ORFs encoding active gag, protease, integrase and a protein termed cORF with features suggestive of the lentiviral Rev protein (Lower et al. 1996; Patience et al. 1997). Interestingly, the normal tissues (non-transformed) with the highest HERV mRNA and protein expression are the placenta and specifically the extraembryonic trophoblast tissue (Wilkinson et al. 1994). Trophoblast cells protect the developing embryo from a macrophage-mediated maternal immunological response. HERV *env* proteins, like those of infectious retroviruses, may possess both immunosuppressive and fusogenic activities (Ruegg et al. 1989; Venables et al. 1995), and have been postulated to mediate immunosuppression of maternal placental macrophages, preventing rejection of the embryo "allograft" (Villareal 1997) which is a long-observed immunological phenomenon which still defies explanation. Surprisingly, the single-copy HERV-R (or ERV3) element encodes an *env* protein which accumulates to very high levels (0.1% of total cell protein) during syncytiotrophoblast differentiation (Venables et al. 1995), and is the only HERV-R ORF maintained over 30 million years of evolution (Patience et al. 1997). Although the conservation and expression data imply a functional role for HERV-R *env* in the trophoblast, the presumed function must be dispensable as recent reports have shown that approximately 1% of Caucasians are homozygous for a mutation causing a premature stop codon (de Parseval and

Heidmann 1998) without apparent effects. The mammalian trophoblast has immunosuppressive and fusogenic properties, but it is not known which, or any, of these functions are mediated by the variety of HERV-R and HERV-K proteins and particles expressed therein (Lower et al. 1996). Interestingly, a trophoblast-specific fusion transcript between a HERV-E element and the growth factor pleiotrophin, when specifically deleted *in vivo*, reverses the invasiveness and tumorigenicity of trophoblast-derived choriocarcinoma cells (Schulte et al. 1996).

While HERV expression is implicated in the normal immunological state of pregnancy, a HERV-K-encoded protein is a causal candidate in an autoimmune disease state and a ERV9-like element is implicated in another. Insulin-dependent diabetes mellitus (IDDM) is an autoimmune disease, caused by a self-reactive T-cell-mediated inflammation and destruction of pancreatic islet β cells (Tisch and McDevitt 1996). Genetic and epidemiological evidence suggests a number of genes are involved in the disease pathogenesis, and that unknown environmental factors likely play a role in triggering the disease onset, and may influence its clinical course (Conrad et al. 1997). A specific subset of T cells positive for the Vβ T cell receptor was discovered to be expanded in two patients with type I IDDM. This preferential expansion of a Vβ -positive T-cell subset is a hallmark of an MMTV-type superantigen (SAG): a protein of viral or bacterial origin which can strongly and reciprocally activate specific subsets of leukocytes involved in the disease state (Huber et al. 1994; Weber et al. 1995). The finding that cultured leukocytes from a subset of IDDM patients released reverse transcriptase activity led to the cloning of a HERV-K-related transcript, by an elegant PCR method (Conrad et al. 1997). This HERV-K was shown, by a variety of functional assays, to encode a SAG function in the N-terminus of the

env protein. Although a SAG activity has been previously demonstrated in a murine autoimmune system, encoded by a small ORF in the 3' LTR of mouse mammary tumor virus (MMTV), this is the first demonstration of a human disease-associated SAG (Choi et al. 1991). A side effect of the expression of MMTV-SAG from an endogenous MMTV retrovirus is the conferral of protection against exogenous MMTV infection (Golovkina et al. 1992). It may be that the HERV-K-SAG was associated with conferring resistance to an ancient HERV-K-like exogenous retrovirus at some point in primate evolution, and now possesses a negative effect in genetically susceptible individuals. It is provocative that, although the location of the SAG moiety differs between MMTV and HERV-K, MMTV is the closest exogenous virus relative of HERV-K (Medstrand 1996). Interestingly, another HERV-K LTR found in the HLA-DQ region has been found to be associated with susceptibility to IDDM (Badenhoop et al. 1996).

Another complex and widespread disease of enigmatic etiology is multiple sclerosis (MS), a neurological disease commonly affecting young adults. While autoimmunity is known to be causal in IDDM, its role in MS is only suggested and still somewhat controversial (Stinissen et al. 1997). Additionally, an infectious agent, perhaps a virus, has been proposed to be involved, with the genetic background of affected individuals playing an important role. The MS virus-association was strengthened by the recent finding of extracellular virion production with RT activity in a MS patient-derived cerebrospinal fluid cell culture (Perron et al. 1991). These virions were purified from various MS patient cell supernatants by sucrose gradient centrifugation, fractionated, treated with nucleases to remove any extracellular RNA/DNA, and lysed. The lysate fractions were divided and RT-activity assays and *pol* region RT-PCR with primers from conserved regions was performed in two

independent laboratories, in a double-blind fashion (Perron et al. 1997). These stringent but necessary controls for spurious amplification, upon final correlation between RT-PCR clones and fraction RT-activity, yielded two types of *pol* sequences, one homologous to the class I HERV, ERV9, and the other, derived from a novel HERV or retrovirus with 75% homology to ERV9, termed MS-associated retrovirus (MSRV). No *pol*-related PCR products were generated from normal control B cells or from RT activity-negative gradient fractions. The MSRV-*pol* RNA was found in cerebrospinal fluid from 5 of 7 untreated MS patients, but not in MS patients treated with immunosuppressives, or in any of 10 patients exhibiting other neurological diseases. Further sequencing of this clone showed that it is highly related to ERV9, and genomic results show it is present in multiple copies in the human genome, as expected of a HERV. Although genomic data was only referred to, and not shown in this study, it appears likely that MSRV is an ERV9-related HERV family with extensive functional protein-coding capacity (Perron et al. 1997). Future genomic cloning and characterization of this HERV is necessary for complete understanding of its role in human disease. Although a causal pathogenic effect of this MSRV in MS remains to be shown, the correlation of expression and disease is provocative and, even if excluded from pathogenesis, may serve as marker of disease state (Garson et al. 1998). It appears that although the major HERV / disease associations have autoimmunity at their basis, much more dissection of cause and effect remains to be done (Andersson et al. 1998).

1.5 HERV effects upon adjacent cellular genes and disease

Potential methods by which ERVs could affect normal expression of adjacent cellular genes have been detailed in the previous section. While many examples are known of both germ line and somatic insertions of murine ERVs causing mutations and malignancies (Fan 1994), no such examples are known for HERVs. This is due to two reasons: the mouse has many more retrotranspositionally active ERVs than the human (approximately 60-fold, as estimated by (Kazazian and Moran 1998)), and for a HERV insertion causing an effect on a nearby gene, that effect has to be beneficial or (more likely) neutral to the host for the insertion to be maintained in the germ line. A somatic HERV insertion activating a oncogene and causing cancer likely kills the host, and detrimental germ line insertion would be selected against if passed on to progeny. Additionally, the precise molecular etiology is rarely examined for sporadic, non-familial cancers. Thus, HERV effects upon cellular genes must be neutral or beneficial to allow their conservation (and subsequent study) in the primate and human germ line. This property makes it difficult to identify such interactions, especially if the effects are neutral or upon an uncharacterized gene, as is most likely. In spite of this handicap, a number of HERV effects upon adjacent genes, mediated by the transcriptional elements in LTRs, have been recently described. These examples are summarized in Table 2. Effects of HERV-H elements upon adjacent genes are presented in the following section.

Table 2 HERV effects upon adjacent human genes

HERV Element	Gene affected	Mechanism	Effect	Reference
HERV-K	HLA-DQB1	Possible enhancer Possible deletion	Disassociation of DQ expression from HLA-DR/DP?	(Kambhu et al. 1990)
HML-6	HLA-DRB6	LTR replaces promoter and 1st exon	New exon	(Mayer et al. 1993)
HML-6	MCIB	Polyadenylation	Downregulates transcription	(Medstrand 1996)
HERV-E	Pleiotrophin	Enhancer	Confers trophoblast-spp. expression	(Schulte et al. 1996)
HERV-E	Amylase	Enhancer	Confers salivary gland spp. expression	(Ting et al. 1992)
ERV9	ZNF80	Promoter	Sole promoter, hemopoietic expression	(Di Cristofano et al. 1995)
HERV-R	plk	Promoter, 5' fusion with env	Activates transcription, possible monocyte-spp. expression	(Kato et al. 1990; Abrink et al. 1998)

1.6 HERV-H and effects upon heterologous genes

The HERV-H family is the most numerous of HERV elements in the human genome, numbering approximately 1000 copies of 5.8 kb unit length and 8.7 full-length elements, and another 1000 copies of solitary LTRs, per haploid human genome (Wilkinson et al. 1994). The majority (~900) of HERV-H elements belong to the 5.8 kb unit length category, being deleted for the entire *env* region and sections of *pol*, but a smaller subset (~100) of elements are essentially undeleted and possess the typical *gag-pol-env* internal structure (Hirose et al. 1993). The internal regions of these elements have accumulated many mutations, with no intact HERV-H ORFs published to date. However, a HERV-H sequence containing an intact *env* ORF has been recently isolated (M. Lindeskog, unpublished observations). HERV-H is a Class I HERV, a member of the group of HERVs sharing homology to mammalian C-type retroviruses. Specifically, HERV-H is most similar to the exogenous murine leukemia virus (MLV)(Mager and Freeman 1987), and is most closely related to the ERV9 family of HERVs (Medstrand 1996). In contrast to the mutated, non-coding internal genes, some of the 400-450 bp LTRs have been shown to possess functional RNAP II promoters, enhancers and polyadenylation signals (Mager 1989; Feuchter and Mager 1990); the promoters apparently require Sp1 binding sites for activity (Sjottem et al. 1996). HERV-H LTRs are grouped into 3 types based on repeats present in the promoter-containing U3 region: Types I, Ia and II. As it contains features of both Type I and II LTRs, Type Ia LTRs likely arose by recombination between the original two types, and concordant with this view, Type Ia LTRs are less numerous in the human genome, likely due to Type Ia LTRs being "younger", or having arisen more recently in

evolution, than Types I or II (Goodchild et al. 1993). Significantly, Type Ia LTRs are normally expressed in a wider range of tissues and are more transcriptionally active than are Types I or II, respectively (Feuchter and Mager 1990; Goodchild et al. 1993). This greater promoter activity has been associated with the gain of binding sites for the Sp1 basal transcription factor and loss of Type I/II repressor sequences (Nelson et al. 1996). The more recent appearance of Type Ia LTRs is an example of the progression of molecular evolution of HERV-H elements in general. HERV-H elements are first detectable in the genomes of New World monkeys, numbering approximately 25-50 elements, with over 50% derived from full-length elements (Mager and Freeman 1995). The deleted form then underwent a large expansion in the Old World primate lineage, about 40 million years ago, to about 900, which is essentially unchanged today. Approximately 20 MYA, HERV-H elements containing the Type Ia LTR then expanded to the approximately 100 copies seen in humans (see Figure 3.5) (Goodchild et al. 1993).

Like most HERVs and LINE elements, endogenous expression of HERV-H can be seen by PCR in most tissues, but is by far the highest in cells derived from early embryonic stages: teratocarcinomas, placenta and the associated chorionic and amniotic membranes (Wilkinson et al. 1990). Interestingly, and also paralleling other HERVs and LINEs, HERV-H transcription is radically reduced when the primitive teratocarcinoma cell lines are induced to differentiate with retinoic acid (Wilkinson et al. 1994). The primary transcript seen is a 5.8 kb unit-length mRNA, with a 3.7 kb spliced transcript being secondary. This smaller transcript was seen to splice out *gag* sequences and retain *pol* regions, and is due to usage of the conserved splice donor signal just upstream of *gag* splicing to a cluster of likely fortuitous splice acceptor sites

near the end of *gag*. This is unique among retroviruses and ERVs, as most subgenomic splicing in exogenous retroviruses use the same splice donor but instead use a single acceptor in the *env* gene, which is lacking in most HERV-H elements (Wilkinson et al. 1994).

The copy number of full length HERV-H elements, exceptionally high among HERVs, make it an ideal candidate to investigate the effects these numerous elements have upon adjacent cellular genes. As shown in Figure 1.2, there are a number of mechanisms by which HERV-H elements can exert influence upon nearby genes, usually at the transcriptional level. To date, influences and associations of HERV-H have been reported for 7 human genes, with several more being currently investigated, and all but 2 originating from the Mager laboratory. Most HERV-H / gene interactions have been discovered as a result of various directed cDNA library screening experiments, designed to detect chimeric transcripts containing HERV-H sequences and unrelated cellular sequences. A primary class of HERV-H / gene influences are composed of 3' fusions of adjacent cellular genes to HERV-H sequences, as exemplified in Figure 1.2C. The first identified and characterized 3' HERV-H / gene fusion was with the PLT gene, isolated from a normal human placental cDNA library as part of a screen for LTR-polyadenylated transcripts (Goodchild et al. 1992). Consisting of a 1.2 kb cDNA containing an anonymous 223 amino acid ORF, this transcript was fused to a HERV-H LTR, which provided the polyadenylation signal. PLT was shown to be a single copy gene of wide expression in human tissues, with additional transcripts possessing variant 3' ends being isolated. Thus, it appears that the HERV-H LTR provides a alternative poly (A) signal for the PLT gene (Goodchild et al. 1992). The PLT gene was recently found to share significant amino acid similarity

with the human Farber disease gene, an acid ceramidase enzyme, to the extent that PLT likely encodes an undescribed type of ceramidase (Koch et al. 1996).

The other partially characterized examples of a HERV-H 3' gene fusion are the HMGIC transcripts derived from various mesenchymal tumors (Kazmierczak et al. 1996). The HMGIC gene is a member of the DNA-binding HMG group of proteins which compose the non-histone component of chromatin. A 3' RACE strategy to detect HMGIC fusions, common in mesenchymal tumors, found the same segment of HERV-H 3' LTR fused to exon 3 of HMGIC, in three different patient tumor samples. As above, the poly (A) signal was provided by the LTR, causing a premature truncation at exon 3 of the HMGIC gene. The resulting proteins would possess the DNA-binding domains and lack the acidic domains found in all normal HMGIC proteins. The lack of additional experimental data in the single published report prevents any further speculation about the role this fusion gene may play in the pathogenesis of mesenchymal tumors (Kazmierczak et al. 1996). The previous two examples of HERV-H polyadenylation of heterologous cellular transcripts were identified by different *in vitro* molecular biological methods. Currently, the most fruitful method of discovering HERV / gene fusions appears to be library screening "*in silico*", i.e. screening the immense expressed sequence tag (EST) databases for cellular sequences fused to HERV sequences (Marra et al. 1998). This method is more useful for identifying 3' fusions or polyadenylations due to the intended and intrinsic 3' bias of the EST sequences (as 3' UTRs are generally highly variable and serve as a "unique" tag for a gene better than coding regions (Schuler et al. 1996)). Database searches (http://ncbi.blm.nih.gov/dbEST) have revealed a variety of apparent HERV-H

polyadenylations of unique transcripts, which are currently under investigation (D. Mager, personal communication).

One example of a 5' HERV-H:gene chimera is the human ZNF177 cDNA, which possesses a small 86 bp segment of a HERV-H *env* gene spliced into the 5' UTR, in reverse orientation (Baban et al. 1996). The fusion was fortuitously discovered during a screen of a NTera2D1 cDNA library for expressed HERV-H *env* sequences. ZNF177 is a KRAB-box type of zinc-finger transcription factor of unknown function, and in addition to the HERV-H sequence, also contains a partial Alu sequence in the 5' UTR, which is another example of the aforementioned clustering of retroelements. The HERV-H sequences are located in a 5' intron of the ZNF177 gene and are incorporated by splicing; interestingly, the element appears to be deleted in a novel manner, with only the *env* and *gag* regions remaining, without LTRs or *pol* sequences (Baban et al. 1996). The 5' UTR is extensively alternatively spliced, with at least 6 different forms being generated in teratocarcinoma mRNA. Expression of this gene is detectable in all tumor cell lines and normal tissue RNAs tested, with transcription appearing low, as the transcript was only detectable by RT-PCR, not northern blotting. This is consistent with the generally low levels of normal expression seen with most zinc-finger transcription factors. ZNF177 is the first example of a gene being expressed from its native promoter subsequently incorporating HERV-H DNA via splicing (Baban et al. 1996). As 5' UTRs are known to regulate translational efficiency (Kozak 1992) it is possible, although currently untested, that the *env*-derived sequences may serve to positively or negatively regulate the translation of this transcription factor.

Another class of characterized HERV-H / gene interactions consists of 5' fusions of HERV-H LTR and internal sequences to adjacent cellular transcripts, with the fusions most often being carried out by splicing, and occasionally by read-through transcription. Derived from a prostate cancer cell line cDNA subtraction library, the hybrid HERV-H / calbindin transcript appears to be promoted by a HERV-H 5' LTR and then spliced to downstream calbindin D28K exons (Liu and Abraham 1991). As the HERV-H sequence ends precisely at the conserved splice donor site, and calbindin sequence starts immediately following, at the start of the second exon, it appears the HERV-H element spliced into the nearby calbindin gene, as diagrammed in Figure 1.2B. The genomic structure of this locus has not been investigated, so the exact position of the HERV-H element with respect to the calbindin exons cannot be established, nor whether the HERV-H element was inserted *de novo* into this locus and was involved in the pathogenesis of the prostate cancer. Interestingly, this fusion transcript contains almost all of the correct calbindin reading frame, with only the 5' end being replaced by 50 amino acids of fortuitous HERV-H-derived ORF, which may be translated in the parental PC3 cells (Liu and Abraham 1991). The fusion removes the first exon of calbindin which abrogates the first of 6 calcium-binding EF-hand motifs. It is unknown what effects this aberrantly promoted, 5' fusion protein may have upon cellular homeostasis, but it is speculated it may contribute to the metastatic phenotype of these cells (Liu and Abraham 1991).

The final example of a HERV-H-promoted cellular gene is similar in structure to HERV-H /calbindin, and is the subject of this thesis. The PLA2L cDNA was cloned as part of a directed search for human cellular transcripts which initiate in a HERV-H LTR and subsequently splice (using the conserved splice donor site, Fig.1.2) into adjacent

genes. This search entailed differential hybridization screening of an NTera2D1 cDNA library to identify clones which hybridize to HERV-H probes directly upstream of the splice donor site, and do not hybridize to HERV-H probes downstream of the site, as expected of unrelated cellular sequences. The primary cDNA clone, termed AF-5, contains the requisite R/U5 regions of a HERV-H LTR at the 5' end, as expected of a LTR-promoted transcript, and also contains directly downstream of the LTR, the HERV-H leader region which ends precisely at the conserved splice donor site (Feuchter-Murthy et al. 1993). The 2.4 kb mRNA is derived from a single copy human gene and is abundantly transcribed in Tera1 teratocarcinoma cells, and to a lesser extent in the cells of origin, NTera2D1. No transcription was seen in a variety of other cell lines and tissues, and no evidence of transcription from a non-LTR promoter was detected. Also, no evidence of a teratocarcinoma-specific rearrangement causing PLA2L and HERV-H fusion was detectable using genomic Southern blotting of 8 normal individuals. The LTR sequence fused to PLA2L was determined to belong to the Type I class, which are expressed at the highest level in Tera1 cells (Wilkinson et al. 1993). Surprisingly, extensive alternative splicing of this gene was detected by northern hybridization and RT-PCR, and several differentially spliced related cDNA clones were isolated (see Fig. 4.9). The spliced chimera contains a 689 amino acid ORF encoded by over 20 exons, with two discrete domains of amino acid level similarity (30-38% identity) to the complete sequence of secreted phospholipase A_2 ($sPLA_2$) (Fig. 3.3). Secreted phospholipase A_2s are a large and ubiquitous family of well characterized small lipolytic enzymes, involved in processes as diverse as lipid metabolism to composing the main toxic component of snake and bee venom (Dennis 1994). The similarity extends over the whole reading frame of $sPLA_2$, and appears to

have selectively conserved the Cys residues important for the highly disulfide-bonded rigid tertiary structure characteristic of $sPLA_2$ (see Table 4 and Fig.4.5). Interestingly, the phospholipase active site is not conserved, and it is therefore unlikely that the PLA2L cDNA could encode phospholipase activity (Feuchter-Murthy et al. 1993). The gene duplication events giving rise to the PLA-domains were necessarily ancient, as dendrograms comparing the domains to functional $sPLA_2$s place the PLA-domains on a separate branch from active, contemporary $sPLA_2$ enzymes. It should be noted that the PLA-domains are discrete polypeptide domains within a larger protein and, although clearly related, are not members of the $sPLA_2$ gene family of lipolytic enzymes. The rest of the large PLA2L ORF lacked any detectable homology to known cDNAs or proteins, at the time of discovery.

Upon characterization of the alternate PLA2L cDNAs, it became apparent that the alternative splicing was localized to the anonymous upstream exons, and the downstream PLA-domain exons were generally spliced as a single transcriptional unit. Interestingly, the 3 additional cDNAs isolated, AF-6, 7 and 8, contained only the upstream exons, including a large 2kb exon with characteristics of a 3' UTR, and lacked the PLA-domains. cDNAs including both upstream exons and PLA-domain exons were only isolated from the cDNA library once, although further clones containing both regions were isolated by RT-PCR on teratocarcinoma mRNA. Surprisingly, the converse situation, evidence of PLA-domain exons being transcribed without the upstream exons, could not be demonstrated (Feuchter-Murthy et al. 1993). Thus, when work on this thesis began, it appeared that the PLA2L transcript encoded a novel human gene promoted by a HERV-H LTR promoter which conferred teratocarcinoma-specific expression. These initial experiments have initiated,

described here, studies into the role HERV-H appears to play in the regulation of the unrelated adjacent locus, PLA2L.

1.7 Rationale and Thesis Objectives

The overall goal of this thesis is to use the elucidation of the influences HERV-H has upon the adjacent gene PLA2L as a paradigm for the effects the family of HERV-H elements has upon the human genome, and by extrapolation, upon human biology. Characterization of the effects a HERV-H element has had upon the PLA2L locus included studies at the level of transcription and translation. The specific questions addressed were:

- What is the molecular nature of the HERV-H / PLA2L fusion?
- What is the genomic structure of the PLA2L locus and where is the HERV-H element relative to the downstream exons?
- When in primate evolution did the HERV-H element integrate into the locus. That is, what species lack the HERV-H promoter for PLA2L?
- Is PLA2L conserved in evolution, and how far have the PLA-domains diverged from functional $sPLA_2$ enzymes?
- Can a PLA2L ortholog which is expressed from a native (non-LTR) promoter be isolated?
- What is the native expression pattern and function of PLA2L?
- Does the HERV-H insertion alter the transcription or translation of the PLA2L locus in novel ways?

The results of this thesis are presented in three chapters. Chapter 3 details the genomic cloning of the PLA2L locus and molecular evolution studies determining the approximate age on HERV-H insertion. The precise intron/exon structure of the PLA-domain region of PLA2L was determined and compared to that of secreted PLA_2. Furthermore, the PLA2L locus was chromosomally and regionally localized and evidence of conservation in a variety of species including mouse, was obtained.

Chapter 4 contains details of the genomic and cDNA cloning of the partial murine homologue of PLA2L, otoconin-90. Studies of this ortholog, combined with other experiments including further human genomic cloning and signal sequence predictions lead to a novel hypothesis for PLA2L biogenesis.

The effects of HERV elements upon adjacent human genes are of great interest; Chapter 5 elucidates the effects that the HERV-H sequences has had upon expression, specifically the translation, of the PLA2L transcript. Fusion proteins and polyclonal anti-PLA2L antibodies were generated, and used in western blotting to examine endogenous and heterologous PLA2L expression. The analysis demonstrates PLA2L-specific HERV-H effects resulting in translational suppression.

CHAPTER TWO : MATERIALS AND METHODS

2.1.1 Library Screening and Genomic Cloning

To construct the PLA2L genomic map, two human genomic libraries were screened with PLA2L cDNA probes. Approximately 1.0 million clones from a human genomic DNA library derived from normal female peripheral blood mononuclear cells and constructed in the λGEM-12 vector were screened by plaque hybridization to Probe 2, a 410 BbsI restriction fragment of the PLA2L AF-5 cDNA (Feuchter-Murthy et al. 1993). An 18.6 kb λ phage clone termed λPLA2L was isolated and found to contain the downstream third of the PLA2L locus, including the PLA-domains. An arrayed human genomic DNA P1 library with a 1.2X coverage of the human genome was obtained from the Reference Library Database (RLDB, (Zehetner and Lehrach 1994)) and screened with a mixture of a 599 bp HincII/ApaI fragment (Probe 1) and Probe 3 from the AF-5 cDNA (Fig. 3.1). A positive clone, P1N1567, was isolated and found to contain the upstream third of the PLA2L genomic locus, including the HERV-H element. Long-range genomic PCR was then performed to close the ~30kb gap between the P1 and λ genomic clones, respectively. All long-range PCR was performed on normal human female genomic DNA, at a concentration of 100-500 ng/reaction. The standard 50 μl long-range PCR reaction contained 10 pmoles each primer, 200 μM of each dNTP, 1 unit Elongase DNA Polymerase mix (Life Technologies), 60 mM Tris-SO_4, 18 mM $(NH_4)_2SO_4$ and 2 mM $MgSO_4$, in a aqueous solution. Primer pairs XPIE/XT7 were used to amplify PLAGAP1 (using a T_m of 59°C), primers P13'End/Gap5'Anti (using a T_m of 63°C) were used to amplify the PLAGAP2 genomic clone, while primer pair Gap5'/Gap3' (using a T_m of 68°C) were used to

amplify the PLAGAP3 clone. The cycling parameters were an initial 30 second incubation at 94 °C, followed by 35 cycles of 94 °C for 30 seconds, the above T_ms for 30 seconds, and 68 °C for 20 minutes. Primer sequences can be seen in Table 3.

To isolate homologous murine genomic clones, a gridded C57BL/6J mouse genomic P1 library was obtained from the Reference Library Database (Zehetner and Lehrach 1994) and screened with a mixture of a 383 bp PstI fragment (residues 995-1378, analogous to Probe 2) and Probe 3 from the human PLA2L cDNA (see Fig. 3.1)(Feuchter-Murthy et al. 1993). These fragments span the two domains of homology with PLA_2. Hybridization was carried out for 16 hours at 55°C in a solution consisting of 7% SDS, 0.5M sodium phosphate, 1 mM EDTA, and was subsequently washed twice for 45 minutes at 55°C in a 40 mM sodium phosphate/0.1% SDS solution. Two positive P1 clones were isolated, P1219 and P1200, with P1219 being ascertained by hybridization to human probes spanning the PLA2L cDNA to contain a larger insert. P1219 was then used in all subsequent experiments. P1219 was digested with various restriction enzymes and then "shotgun" ligated into Bluescript (Stratagene). This ligation mixture was transformed into DH5α competent *E.coli* and clones were picked onto a gridded plate and screened by colony hybridization. A 5.8 kb HindIII fragment (308H) was identified by hybridization (at reduced stringency, 55°C) to a human 308 PstI PLA2L cDNA fragment (Probe 5). Using a mouse genomic probe derived from the 3' end of 308H (a 651 bp AccI fragment), an overlapping 4.1 kb BamHI fragment was cloned. The 4.1 kb DNA fragment also hybridized to the 308 PstI probe. Subclones from both the 5.8 kb and 4.1 kb fragments were partially sequenced to confirm the presence of PLA2L-homologous exons.

2.1.2 Evolutionary Genomic PCR and Chromosomal Mapping

PCR reactions were performed on genomic DNAs using a standard 50ul reaction containing 100ng genomic DNA, 30pmoles of each primer, 250 uM of each dNTP, 1X PCR buffer (Life Technologies PCR Buffer containing 20 mM Tris-HCl, pH 8.4, 50 mM KCl), 1.5 mM $MgCl_2$ and 1.25 units Taq DNA Polymerase (Life Technologies). PCR was performed in a Ericomp TwinBlock thermal cycler using one cycle of 2 minutes at 95°C, 30 cycles of 1 minute each at 94, 48 and 72°C followed by 5 minutes at 72°C for primer pair A1 and A2. The cycling parameters for primer set B1 and B2 (Table 3) were one cycle of 2 minutes at 95°C, 30 cycles of 1 minute at 94, 59 and 72°C, followed by 5 minutes at 72°C. Chromosomal mapping was performed by PCR, using as a template DNA from the NIGMS human/rodent somatic cell hybrid mapping panel 2 (Corielle Cell Repository, Camden, NJ) (Drwinga et al. 1993). Primer set PLAMAP/S42A was used to amplify a 585 bp genomic fragment of PLA2L. The PLAMAP/S42A primer set is derived from exons A4 and B, respectively (Fig. 3.2). Cycling parameters for this primer set were 2 minutes at 95°C, 30 cycles of 1 minute at 94, 57 and 68°C, followed by 5 minutes at 72°C. All sets of PCR reactions were electrophoresed on a 1.2% TBE-agarose gel and stained with ethidium bromide. Following chromosomal assignment to human chromosome 8, regional mapping was carried out using the above PCR conditions on a panel of somatic cell hybrids carrying fragments of chromosome 8 (Wagner et al. 1991; Wood et al. 1992).

Table 3 PCR Primers Used

Primer Name	Sequence (5'- 3')	T_m (°C)
A1	AAACTAATATCTGAGCCCCACTTCCTTCTTTT	84
A2	CTGAAAGGTCACTGGACTGC	62
B1	TGTCAGGCCTCTGAGCCCAAGC	64
B2	GCCCTCAGCCTCTCCAG	58
PLAMAP	GTGCTGATCCAGTTTGTCAA	58
S42A	CCGACACAATCGACCTC	54
XPIE	ACTACGCGTGAAGAGCCAGTCCTGGTCC	68
XT7	GAATTGTAATACGACTCACTATAGG	68
P13'END	TAGGTACCGACACAGCCTGCACTGAGAACTC	72
GAP5'ANTI	CAGGTACCCTTCATCAGAACAGATACAGTCACT	70
GAP5'	CAGGTACCAGTGACTGTATCTGTTCTGATGAAGGA	76
GAP3'	ACGGTACCACTGGTGAGAGGAATAACAACAG	76
MEST3	CCATGCTCTGGACACACCAAAT	66
MEST4	ACTCGTCCACAGGCATCCCTT	66
MEST5	TTTTTGACAGTCCTGGAGGCAA	64

2.1.3 Cell Lines and Nucleic Acid /Protein Extractions

Primate cell lines KG1a (human), Wes (chimpanzee), ROK (gorilla), Puti (orangutan), MLA144 (gibbon), 26CB-1 (baboon), CV-1 and COS7 (African green monkey), and B95-8 (marmoset) were obtained from the American Type Culture Collection and grown in culture as recommended. The cell line Tera1 is a human

teratocarcinoma cell line derived from testicular germ cell carcinoma (Fogh and Trempe 1975) and was obtained from the ATCC and grown as recommended. Slow loris DNA was kindly provided by Dr. Morris Goodman and dog DNA was obtained from Dr. Paula Henthorn. Genomic DNA was isolated using established protocols (Sambrook et al. 1989).

Total RNAs were extracted from cell lines using Trizol (Life Technologies), following the manufacturer's protocol.

Protein lysates from cell lines were prepared by washing the pelleted cells twice in phosphate-buffered saline (PBS) then solubilized with 0.5% Nonidet P-40 in phosphorylation solubilization buffer (PSB) for 1.5 hours (Liu et al. 1994) at 4°C. PSB is composed of 50 mM N-2-hydroxyethylpiperazine-N'-2-ethansulfonic acid (HEPES, pH 7.4), 100 mM NaF, 10 mM $NaPP_i$, 2 mM Na_3VO_4, 4 mM EDTA, 2 mM PMSF, 10 μg/ml leupeptin and 2 μg/ml of aprotinin. Nuclei were then cleared from the lysate by a brief centrifugation, and the cleared lysate was stored at -20°C. Cell lysate from the murine hemopoietic cell line BAF3 was kindly provided by Mark Ware.

2.1.4 Southern and northern blotting and hybridizations

Genomic Southern analysis was performed by electrophoretically separating restriction enzyme digested genomic DNAs on 0.8% TAE-agarose gels, and transferring the DNA to Zetaprobe (Bio-Rad) membranes via the alkaline blotting method. Blots were hybridized using the buffer of (Vanin et al. 1983) altering only the final concentration of SDS to 2%. The "zooblot" genomic Southern (Fig. 3.7) was hybridized to a 318 bp HindIII/AvaI fragment of the PLA2L cDNA (Probe 3, Fig. 3.1), with the final post-hybridization wash being performed at 60°C in 1X SSC. Southern

analysis of primate genomic PCR products was performed using 61°C hybridization to Probe 4 (Fig. 3.1), a 98 bp Nde I/AlwNI PLA2L cDNA restriction fragment, followed by a 61°C wash in 3X SSC, 1% SDS, for 1 hour. 1X SSC is 0.15 M NaCl / 0.015 M sodium citrate.

Total RNAs were fractionated on 1% formaldehyde-agarose gels and northern blotted onto Zetaprobe (Bio-Rad) membranes in 10X SSC. The membranes were fixed by UV crosslinking, and hybridized in freshly made 50% deionized formamide, 5% SDS, 0.5M NaH_2PO_4 in 1 mM EDTA pH 7.2, and 1 mg/ml bovine serum albumin (Sigma). All northern solutions were made with diethylpyrocarbonate (BDH) treated, autoclaved Milli-Q water. Prehybridization was carried out for 1 hour at 42°C, followed by hybridization of ^{32}P-labelled DNA probes for 16-20 hours at 42°C. Two initial washes, each lasting 20 minutes, were carried out at 55°C, in 2X SSPE/0.3% SDS (1X SSPE is 0.18 M NaCl, 20 mM NaH_2PO_4, 1 mM EDTA, pH 7.5), followed by two 20 minute washes at 55°C in 1X SSPE/0.5% SDS. The final two 20 minute washes were performed at 60°C in 0.3X SSPE/1.0% SDS, subsequent to which the northern blot was wrapped in Saran Wrap and exposed to X-ray film for 48 hours at -70°C. The northern blot was first hybridized to Probe 2, a 410 BbsI fragment of the PLA-domain of PLA2L (Fig. 3.1), then to a 1.9 kb PstI fragment containing most of the chicken β-actin cDNA. All probes were labeled by the random primer extension method (Sambrook et al. 1989) with α-^{32}P-dCTP (Amersham), at a concentration of 2-3X10^6 dpm/ml for probes used in genomic Southern and northern hybridizations, and at 2-3X10^5 dpm/ml for non-genomic DNA Southern hybridizations.

2.1.5 Murine cDNA synthesis and RT-PCR

To identify a murine cDNA homologous to PLA2L, a BLAST search was performed on the mouse expressed sequence tag (EST) Genbank database. Three EST clones were found to be similar to the downstream half of PLA2L, and were isolated from a cDNA library derived from pooled E13 and E14 mouse embryos. When compared, these clones were seen to be lacking the 5' end of the mouse homologue, termed otoconin-90, as well as some small gaps in the 3' end. Murine embryo E14 heads were used as a source of cDNA to attempt to clone the complete transcript. Mouse E14 embryos (kindly provided by Dr. Cheryl Helgason) were rinsed in PBS and the heads were dissected. Six heads were pooled and total cellular RNA was extracted using Trizol, during homogenization in a Dounce homogenizer. First strand cDNA was synthesized using both random hexamers and a primer, MEST4, with the primer-directed cDNA being subsequently used. Five μg of total embryo RNA, which had been previously treated with RNase-free DNase I (Life Technologies), and 5pmol of primer were heat denatured at 70°C for 10 minutes, then centrifuged briefly in a microfuge at 4°C. All subsequent steps were carried out using Life Technologies enzymes and buffers. The reverse transcription reaction was then carried out in 1X First strand buffer (50 mM Tris-HCl, pH 8.3, 75 mM KCl, 3 mM $MgCl_2$), 10 mM DTT, 0.5 mM dNTPs and 10 U Placental RNase Inhibitor. This reaction mix was then warmed to 42°C for 2 minutes, and 2 U of Superscript II reverse transcriptase was added, and synthesis was allowed to occur for 2 hours at 42°C. Subsequent to the reverse transcriptase being heat-denatured at 70°C for 10 minutes, one-tenth of the first strand cDNA reaction was used as template for PCR. Negative controls lacking reverse transcriptase were also generated. To close the gap between a 5' exon

sequenced from genomic DNA and the 5' end of the EST clones, PCR primers MEST3 and MEST4, which were derived from the 3' and 5' termini of these DNAs, respectively, were used to perform PCR on the mouse embryo cDNA. Standard cycling parameters were used, with a T_m of 56°C, and a product of the approximate expected size of 280 bp was amplified, cloned and sequenced (Fig. 4.2).

2.1.6 Construction of full length murine otoconin-90 cDNA

The task of generating a full-length cDNA was considerably simplified by the sequencing and deposition of 3 otoconin-90 ESTs by the Washington University Mouse EST Project, shortly following the initial cloning of the murine otoconin-90 genomic locus on P1219. These three EST clones; W50767, AA034721 and AA437511, were all derived from mouse embryo E13-14 cDNA libraries, and did not overlap with the original 5' exon sequenced from mouse genomic clone 308H. This exon, likely the third, was known to be near the 5' terminus of the otoconin-90 cDNA as it encoded the N-terminus of the mature protein. Two EST clone sequence reads, W50767, and AA034721, overlapped by approximately 60 bases (Fig. 4.2). As compared to the human PLA2L sequence, approximately 50 bases were missing between the 3' end of the W50767 read and the 3' EST read, AA437511, which was subcloned from the actual EST clone W50767 (obtained from Research Genetics) as a 293 bp PvuII / PstI fragment, and sequenced. Upon end-sequencing, this W50767 clone also yielded the 53 bases of extreme 3' terminus of otoconin-90, including the polyadenylation signal and poly(A) tail. The 226 bp gap between the initial genomic exon and the 5' end of EST sequence AA034721 was amplified from E14 embryo head RNA with RT-PCR as previously mentioned. Additionally, prior to the sequencing

of EST AA034721, an exon was cloned from P1219 genomic DNA in the pSPL3 vector, using the exon amplification procedure (Church et al. 1994). This procedure allows the isolation of exons from uncharacterized genomic DNAs by virtue of their being flanked by splice donor and acceptor sites. To identify the extreme 5' terminus of the transcript, various 5' rapid amplification of cDNA ends (5' RACE) protocols were attempted upon both random-primed and MEST4-primed embryo head RNA, to no success. Semi-nested PCR upon a E14.5 embryo cDNA library (Novagen) using vector and otoconin-90 primers was then successfully performed. As the library was directionally constructed in the vector λExLox, PCR was initially performed using the vector primer T7Gene10 and MEST4, with a T_m of 59°C. As template, 5 pools of 1:1250 dilutions of the phage library were used, employing an initial 3 minutes at 95°C step, to lyse the phage and denature the library DNA. A 1:100 dilution of the above PCR reactions were used as template in the nested PCR, usingT7Gene10 again as the vector primer, and MEST5 as the otoconin-90 primer, with a T_m of 58°C. These PCR reactions yielded two bands, 250 and 280 bp, which hybridized to the internal oligo, MEST3. Upon cloning in pGEM-T (Promega) and sequencing, the 250bp product was seen to encode 133 bases of novel 5' sequence, including a potential initiating methionine with a strong Kozak consensus match followed by a predicted strong secretion signal sequence (Fig. 4.7).

2.1.7 Plasmid DNA isolation and sequencing

DNA from positive phage clones was isolated using standard procedures (Sambrook et al. 1989) and the inserts subcloned into Bluescript (Stratagene). Plasmid DNA was prepared using a modification of the alkaline lysis miniprep method

(Sambrook et al. 1989). DNA from P1 clone N1567 and P1219 was isolated using a modified alkaline lysis / CsCl gradient procedure (J.Schmidt, personal communication), which entailed steps to minimize shearing of the large P1 clone DNAs. Restriction fragments were subcloned into Bluescript. Primer walking and the exonuclease III digestion protocol were used to sequence all genomic clones. DNA sequence was determined by the dideoxynucleotide chain-termination method or on an Applied BioSystems model 310 fluorescent sequencer with BigDye chemistry (Perkin-Elmer). Alignment and analysis of DNA sequences were performed using GCG software (Devereux et al. 1984) and NCBI BLAST database searches.

2.1.8 Expression of PLA2L:GST fusion proteins and generation of anti-PLA2L antiserum

In order to produce PLA2L antiserum, an immunogenic portion of the PLA2L cDNA was cloned in-frame to glutathione-S-transferase in the vector pGEX2T (Pharmacia). An immunogenic region of PLA2L (nt 398-591) was predicted using the GCG PlotStructure program, amplified using PCR and cloned into the SmaI site of pGEX2T, and sequenced to confirm an in-frame fusion. This construct created a 64 amino acid C-terminal fusion of PLA2L sequences to parasite glutathione-S-transferase under the control of a *tac* promoter. A culture of this construct was grown to log phase, then fusion protein expression was induced by fresh 70 μM IPTG for 5 hours. The bacteria were lysed by sonication and then clarified by centrifugation. The supernatant was then incubated with glutathione-agarose beads (Sigma) in 1 mM PMSF (Boerhinger-Mannheim) while rocking, overnight at 4°C. The beads with bound GST:PLA2L fusion protein were washed twice in 1% Nonidet P-40 (Sigma) in

phosphate-buffered saline (PBS), and stored at 4°C in PBS containing 0.04% sodium azide. To further purify the fusion protein, a fraction of the beads was incubated with an equal volume of 15 mM reduced glutathione (Sigma) to compete the fusion protein away from the glutathione-agarose beads. Aliquots of both the eluted fusion protein and the fusion protein bound to beads were electrophoresed on a 12% SDS polyacrylamide minigel, fixed with 25% isopropanol/10% acetic acid, and stained with Coomassie Brilliant Blue R-250 (Bio-Rad), where the expected 38 kDa, largely pure fusion protein was seen (Fig. 5.2). As GST-fusion proteins bound to agarose beads have been shown to be potent immunogens (Oettinger et al. 1992), and a significant fraction of fusion protein is lost upon competing (Fig. 5.2), the fraction containing bead-bound fusion protein was mixed in a 1:2.5 ratio with Freund's incomplete adjuvant and injected subcutaneously into New Zealand Giant White rabbits (UBC Animal Care Centre). Following a standard boosting regimen, the rabbit was exsanguinated and the blood allowed to clot at 4°C, overnight. The serum was then separated with a 3300 rpm centrifugation and frozen at -70°C. A 1:750 dilution of this polyclonal antiserum was then found to be optimal for ECL-visualized western blots.

An additional PLA2L:GST fusion protein and polyclonal antisera was produced, from a downstream region of the PLA2L transcript, termed pGEX-AF3. This construct was made by PCR amplifying bases 753-959, with a EcoRI site engineered in the 5' end, and cloned into EcoRI/EcoRV-digested pGEX3T (Pharmacia). This fusion protein was expressed exactly as above, but electrophoresed upon a 10% SDS-PAGE, and stained with 0.5 μg/ml ethidium bromide in water for 5 minutes. The gel was then visualized and photographed on a standard UV transilluminator (Fig. 5.2). Polyclonal antisera was generated as above, but not used in the following experiments.

2.1.9 Western blotting and probing

Between 10-30 μg of cell line lysate proteins, as determined by the Bradford assay (Bio-Rad) were electrophoresed on 10%-12% SDS polyacrylamide gels. Proteins were then electrophoretically transferred to Immobilon PVDF membranes (Millipore) using 500 mA for 90 minutes at 23°C in a submarine apparatus containing western transfer buffer composed of 25 mM Tris, 192 mM glycine, 0.05% SDS and 20% methanol. Blots were first blocked overnight at 4°C in a solution of 5% skim milk powder in PBS. Blots were then washed three times, for one hour each, in TBST (10 mM Tris-HCl, pH 7.4, 150 mM NaCl, 0.05% Tween-20) and hybridized to a 1:750 dilution of the anti-PLA2L rabbit polyclonal antiserum in blocking solution for 1 hour. All washing and hybridization steps, unless noted, were carried out at room temperature with constant agitation. The western blot was then washed four times 30 minutes, in TBST. The secondary antibody consisted of horseradish peroxidase-conjugated AffiniPure goat anti-rabbit IgG (Jackson Immunoresearch), which was used at an approximate 1:8000 dilution in TBST for 50 minutes. The western was subsequently washed four times 30 minutes each in TBST, and visualized using the Renaissance ECL kit (NEN-Dupont) and Kodak X-Omat AR X-ray film.

2.1.10 Transfections

Miniprep plasmid DNAs were transiently transfected into 70% confluent COS7 cells using a modified protocol of (Hammarskjold et al. 1986). Briefly, approximately 10 μg of plasmid DNA was mixed with 1 mg/ml of DEAE-dextran (Sigma) and applied directly to the cells for 10 minutes at room temperature, followed by 40 minutes at 37°C. The DNA/DEAE-dextran solution was then replaced by 20% glycerol and swirled

for 2 minutes. The cells were subsequently rinsed in DMEM media (Stemcell Technologies) and DMEM containing 5% fetal calf serum (Life Technologies) and 200 μM chloroquine diphosphate was added for 3 hours, at 37°C. The media was removed and replaced with fresh DMEM/5% fetal calf serum, and the cells were grown for 48 hours and subsequently lysed.

2.1.11 Construction of PLA2L Expression Vectors

To assay PLA2L expression in a heterologous system, two sets of mammalian expression vectors were constructed, for transient expression in COS7 cells. The mammalian expression vector pCDNA3 (Invitrogen) containing the cytomegalovirus early promoter/enhancer and the SV40 origin of replication was used as a backbone for two PLA2L constructs. To construct pPLA2L-Full, the complete 2.4 kb AF-5 cDNA containing HERV-H sequences at the 5' end (Fig. 5.1) was cloned into the EcoRI site of pCDNA3. As our laboratory has previously noticed occasional recombination and instability of HERV-H sequences within the *E. coli* strain DH5α, the ligation mixture was used to transform XL2-Blue (Clontech) and STBL2 (Life Technologies), two independent *E.coli* strains known to suppress some recombinations. The constructs derived from XL2-Blue and STBL2 bacteria were termed pPLA2L-Full1 and -Full2, respectively. A 5' deletion construct, lacking all HERV-H-derived sequences and termed pPLA2L-del, was generated by inserting the 2166 bp Hinc II fragment of the PLA2L cDNA in pBluescript into the EcoRV site of pCDNA3. This construct lacks the 5' 289 bp of sequence found in the complete PLA2L cDNA clone, and deletes all HERV-H derived sequences (1-259) and 30 bp of unique PLA2L sequence. The

vector/insert junctions of all constructs were subsequently sequenced to confirm correct orientation relative to the CMV promoter.

In order to determine whether the HERV-H sequences present in PLA2L can suppress translation of a heterologous gene, two sets of expression constructs were generated with the HERV-H sequences acting as a 5' UTR for the human Thy-1/CD90 gene, a hemopoietic cell surface protein (Craig et al. 1993). The 289 bp EcoRI/Hinc II fragment of PLA2L containing HERV-H sequences was cloned directly upstream of the start codon of the human Thy-1 gene, which was excised from the vector pCTV83 as a 910 bp blunted-MluI/ClaI fragment. This cassette was cloned into EcoRI / ClaI – digested vector pAX142, and named pOD1. pAX142 is a mammalian expression vector containing the strong human elongation factor 1α promoter/enhancer and SV40 origin of replication. Both pAX142 and the human Thy-1 cDNA were kindly provided by Dr. Rob Kay. A positive control containing only the Thy-1 cDNA in pAX142 was also constructed, termed pOD2. As the initial PLA2L translational suppression experiments were carried out using the pcDNA3 vector, an analogous set of HERV-H/Thy-1 chimeras were constructed using pcDNA3 as a backbone vector, to control for any differences in transcription and translation from the different heterologous promoters. Briefly, the Thy-1 cDNA was cloned into the NotI/ApaI –digested pCDNA3, and a 317 bp BamHI/HincII fragment of pPLA2L-FullI containing the HERV-H sequences was subsequently cloned into the BamHI/EcoRV site. The construct containing only Thy-1 was termed pcDNA3Thy1, while the construct using HERV-H sequences as a 5' UTR for Thy-1 was named pcDNA3Thy1HERV.

2.1.12 Anti-Thy-1 Flow Cytometry

To determine the effects the PLA2L-derived HERV-H LTR and internal sequences had upon translation of the heterologous cell surface molecule Thy-1, HERV-H/Thy-1 expression constructs were transiently transfected with DEAE-dextran into COS7 cells. Forty-eight hours post-transfection, the monolayer of transfectants was gently raised and disassociated with PBS/ 1 mM EDTA, 5 minutes at room temperature. All following Thy-1 transfectant manipulations were performed in polystyrene tubes, on wet ice, and covered with an opaque lid. Pelleted cells were resuspended in labeling buffer (Hank's Salt containing 2% fetal calf serum and 0.05% sodium azide). A 1:200 dilution of an anti-Thy-1 monoclonal antibody (5E10), labeled with the fluorochrome phycoerythrin was added, and incubated for 30 minutes, mixing twice by tapping the tube. 5E10 was kindly supplied by Dr. Peter Lansdorp. The labeled cells were then washed twice in labeling buffer, and resuspended by slow vortexing, in 1 mL labeling buffer, and loaded onto a Becton-Dickinson FACScan cell sorter and analyzed with the PC-LYSYS II program (Fig. 5.5). Expression levels of phycoerythrin-labeled Thy-1 molecules on transfected COS cells were acquired using the FL2 channel, and visualized on a 2-axis scatter plot composed of FL2 (phycoerythrin excitation wavelength) fluorescence versus side scatter.

CHAPTER THREE : GENOMIC STRUCTURE AND EVOLUTION OF THE HUMAN PLA2L LOCUS

A majority of the data presented in this chapter composed the following manuscript:

Kowalski P.E., Freeman J.D., Nelson D.T. and Mager D.L. Genomic structure and evolution of a novel gene with duplicated phospholipase A_2-like domains. Genomics 39: 38-46, 1997.

3.1 Introduction

In a previous study, our laboratory reported an association between a novel PLA_2-like transcript, termed PLA2L, and a human endogenous retrovirus (Feuchter-Murthy et al. 1993). Human endogenous retroviruses are most probably derived from ancient exogenous retroviruses, and many families of these sequences exist throughout the human genome (for reviews see Wilkinson et al., 1994; Lower et al., 1996). Several cases have been reported in which a HERV sequence affects expression of an adjacent normal cellular gene (Liu and Abraham 1991; Ting et al. 1992; Di Cristofano et al. 1995). In the case of the PLA2L locus, a retroviral element from the HERV-H family (formerly RTVL-H) has inserted into an intron upstream of the PLA2L exons. Subsequent to the integration event, promoter and enhancer elements in the LTR of the HERV-H element have apparently assumed control of PLA2L expression from the native promoter (Feuchter-Murthy et al. 1993). Structurally, the human PLA2L locus contains several novel features: the presence of, and subsequent promotion by, an endogenous retrovirus-like element, and the presence of duplicated domains of homology to secreted phospholipase A_2, within a much larger open reading frame. The original PLA2L cDNA was isolated from a human NTera2D1 teratocarcinoma cell cDNA library using a subtractive hybridization strategy to identify transcripts promoted by LTRs of HERV-H elements (Feuchter-Murthy et al. 1993). Transcripts corresponding to this cDNA were shown to be present at a low level in NTera2D1 cells, and at a substantially higher level in an independent human teratocarcinoma cell line, Tera1, and not seen in any other human cell line or tissue (+50) examined. The PLA2L cDNA is a product of splicing between retroviral and

cellular sequences and consists of a 2067 bp open reading frame, containing two 360 bp domains of 30 - 38% amino-acid level identity to the complete mature secreted forms of PLA_2. The remainder of the coding sequence is unrelated to any human sequence or EST in the databases. It should be noted that no part of PLA2L is represented as a human EST, implying an extremely restricted pattern of expression, both temporally and spatially.

In terms of structure, activity, cellular roles and expression, phospholipase A_2 (PLA_2) enzymes are among the most-characterized eukaryotic proteins (Dennis 1994). PLA_2s participate in such diverse roles as snake venom toxicity, lipid digestion, and the initiation of inflammation. The enzymatic catalysis effected by PLA_2 is the hydrolysis of the *sn*-2 fatty acid acyl ester bond of phospholipids, which are often cell membrane-associated, liberating free fatty acid and lysophospholipids. PLA_2 enzymes are divided into two primary classes, a 13-15 kDa secreted form, characterized by a rigid three-dimensional structure composed of disulfide bridges and by catalytic dependence upon Ca^{2+}, and a 85 kDa cytoplasmic form (Mukherjee et al. 1994). The many diverse types of secreted PLA_2s ($sPLA_2$) are further differentiated by classifying $sPLA_2$s found in the mammalian pancreas and in elapid venom as Group I, and $sPLA_2$ found in viperid venom and in human inflammatory fluids as Group II, as well as classifying enzymes by the position of cysteine residues (Dennis 1994). Certain amino acids are conserved among and between the Groups, including residues composing the Ca^{2+}-dependent active site and 10-14 conserved cysteine residues required for secondary structure. In addition to canonical Group I and II $sPLA_2$s, many other PLA_2 enzymatic activities exist in various tissues for which the cognate gene or protein has not yet been isolated (e.g. see Ackermann and Dennis, 1995). Further examples of the

complexity of the PLA_2 family of enzymes are: the abundance of inactive but modulatory PLA_2s in certain snake venoms (Maraganore and Heinrikson 1986), and the existence of a cluster of PLA_2-related genes on human chromosome 1 (Tischfield et al. 1996). These findings indicate that the superfamily of PLA_2 homologous genes is both large and complex.

Here I report the cloning of the genomic locus of PLA2L including regions upstream of the HERV-H element. The intron/exon structure of the domains of $sPLA_2$ homology was determined and compared to the structure of other secreted PLA_2s. Using genomic PCR on DNA from various primate species, it was determined that the HERV-H element in the PLA2L locus integrated 15-20 million years ago (MYA). In addition, the gene was chromosomally and regionally mapped to human 8q24.1-8q24.3, and evidence for sequence conservation in other mammals was obtained.

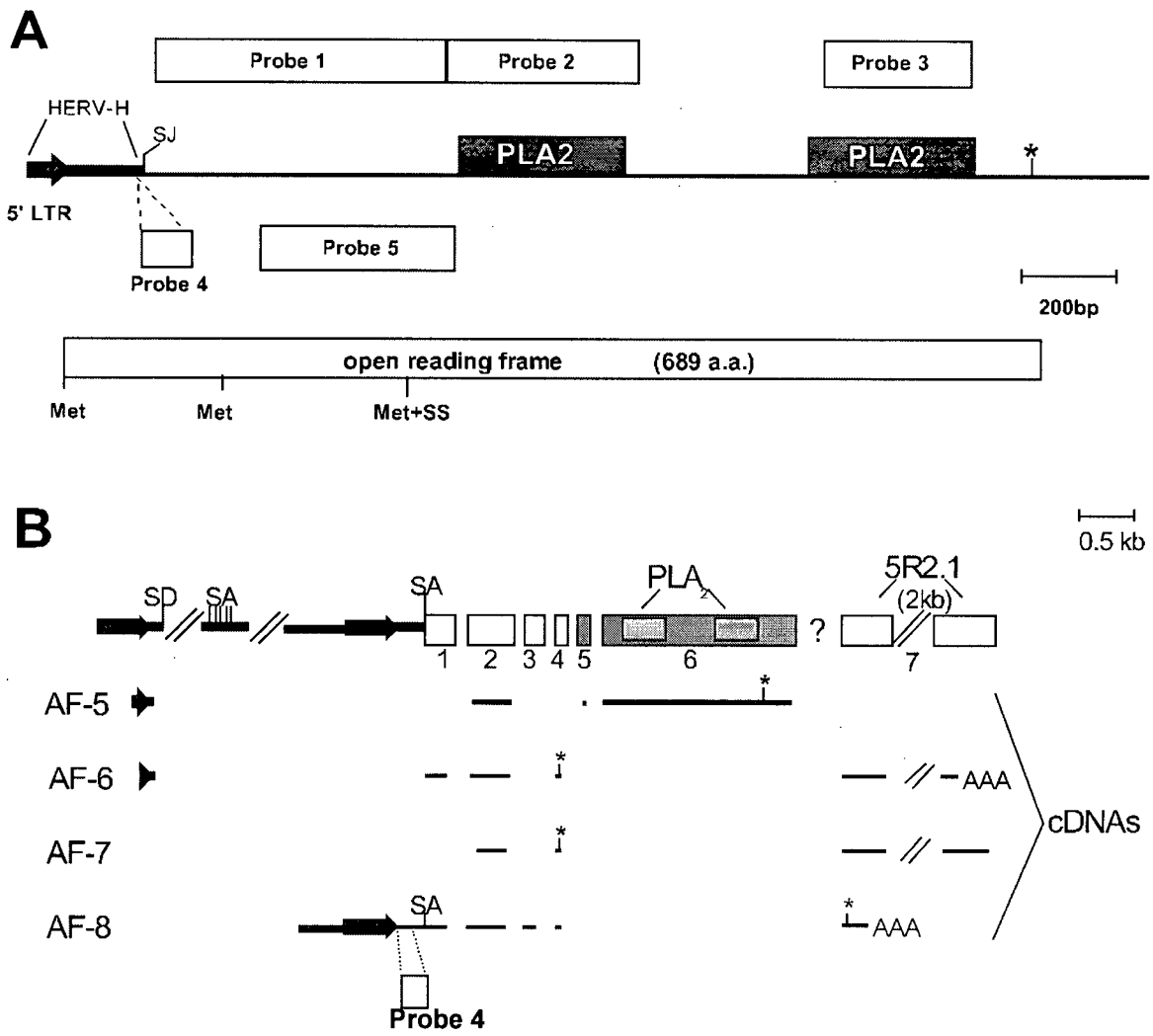


Figure 3.1 Schematic structures of the PLA2L cDNA and probes.
A) The domains of sPLA$_2$ homology are shown as black boxes. The stop codon is indicated with an asterisk. The open reading frame is shown below with the putative initiating methionines indicated. The probes used in this study are indicated as white boxes. Probe 4 is derived from another cDNA, AF-8 (Feuchter-Murthy et al. 1993), and is located directly downstream of the HERV-H 3' LTR (see part B). While not present in the original cDNA clone AF-5 due to alternative splicing, Probe 4 is the most proximal probe to the HERV-H integration site. SJ: splice-junction site between the HERV-H element and PLA2L, Met+SS, putative initiating methionine followed by predicted secretion signal sequence (see Chapter 4). B) Original composite structure of the PLA2L transcript and locus at top, as deduced by comparing and sequencing the original 4 cDNA clones (AF-#, below) and other RT-PCR clones, modified from Feuchter-Murthy et al. 1993. The HERV-H element is shown in thick black lines, and groups of exons are shown as white or gray boxes, and are numbered beneath. Gray exons correspond to ones appearing to be PLA-domain-containing exons which were always spliced together. *; stop codon, paired diagonal lines show a gap, removed for scaling purposes.

3.2 Results

3.2.1 PLA2L Genomic Cloning

Figure 3.1A and B illustrates the 2.4-kb PLA2L cDNA AF-5 that was isolated previously (Feuchter-Murthy et al. 1993). To characterize the PLA2L genomic locus, a human genomic bacteriophage library was screened with two restriction fragment probes, Probe 4 and Probe 2 (Fig. 3.1A). Three clones were isolated using Probe 2, while no clones were isolated using Probe 4, which is located just 3' to the HERV-H element in genomic DNA. I was also unsuccessful in obtaining positives with Probe 4 when an independent bacteriophage library was screened. This result was not totally unexpected as previous observations have shown that HERV-H-containing lambda phage clones can be unstable and very difficult to isolate (Mager and Henthorn 1984). I therefore screened the RLDB human GM1416B lymphoblastoid cell genomic P1 bacteriophage library (Zehetner and Lehrach 1994) and was successful in isolating one clone (P1 N1567) using a mixture of Probes 1 and 3 (Fig. 3.1A). This P1 phage clone spans regions both 5' and 3' to the retroviral insertion. P1 phage do not appear to be immune from HERV-H-induced recombination, however, as this clone was also unstable and it was therefore impossible to measure the insert size.

At least two different rearrangements of the P1 clone were observed, one involving recombination between the HERV-H LTRs as observed previously in lambda bacteriophage (Mager and Henthorn, 1984; unpublished observations) and one involving more complex recombination events. Therefore, all subclones isolated from this P1 were verified by size comparison to human genomic Southern data (data not shown). Overlapping subclones of the region upstream of the HERV-H element were

obtained and the retroviral insertion site verified by identifying the characteristic 5-bp direct repeat flanking all HERV-H integrations (Mager and Henthorn 1984). As the P1 and λ bacteriophage clones did not overlap, I attempted to close the approximate 30 kb gap by performing long-range PCR with cDNA-derived primers, using human genomic DNA as a template. Clones pA4 and pBY2, immediately upstream of the λPLA2L clone, were isolated in this manner. At this point, it was realized that the large 3' regions of the original AF-6 and AF-7 NTera2D1 cDNA clones (known as 5R2.1), which were initially thought to lie *downstream* of the PLA-domain region of PLA2L (Figure 3.1B), were actually found *upstream* of the PLA-domains, in the genomic gap region previously mentioned. This was shown by hybridization, and resulted in a paradigm shift in the view of both the genomic organization of the PLA2L locus and the mechanistic view of the PLA2L transcript's biogenesis, to be discussed further in the subsequent chapter. Using a primer from the 3' end of the AF-5 cDNA clone and a primer from the 5' end of the pA4 genomic clone, long-range genomic PCR was successfully performed, and the 14.5 kb PLAGAP3 clone was generated. This clone encompassed almost half the genomic gap, and contains what is likely the second coding exon, A2, of the downstream PLA-domain containing region. The 3' end of the P1N1567 clone was isolated using long-range PCR with the XPIE primer and a P1 vector specific XT7 primer, to amplify and clone a 10 kb product, called PLAGAP1. Using genomic sequences derived from this clone, the P13'END primer was made, and used in concert with the GAP5'ANTI oligo (which is the antisense to the 5' end of the PLAGAP3 clone), to amplify and clone PLAGAP2. The 5 kb PLAGAP2 clone contains the AF6/AF7 3' cDNA sequences, which are thought to be noncoding 3' UTR sequences. Approximately 2.1 kb of "cDNA" sequence is contained upon a 3.6 kb *Kpn*I

genomic fragment. The 14.5kb PLAGAP3 clone was generated using primers from the 5R2.1 exon and the 5' end of the pA4 genomic clone. All primers salient to the above genomic cloning experiments are shown in Figure 3.2, below panel B, and in Table 3.

3.2.2 Intron-Exon Structure

A genomic map of the region is illustrated in Figure 3.2. The 5' boundary of the P1 clone is not known but extends at least 10 kb upstream of the HERV-H element. The 5' end of the P1 insert was isolated using hybridization to flanking unique vector sequences. Approximately 500 bp of this clone was sequenced, and was not homologous to any known sequences. The complete intron-exon structure was determined for the 5' and 3' domains of $sPLA_2$ similarity including the interdomain region (Table 4), and three exons upstream of the first PLA-domain which belong to the same transcriptional unit. Exons in this region have been denoted A2 - K in Table 4 and Figure 3.2. The regions between the PLA-domains and the HERV-H element contain at least 8 small exons, many of which undergo alternative splicing (Feuchter-Murthy et al. 1993), and the larger 5R2.1 region. The precise locations of these upstream exons has not been determined.

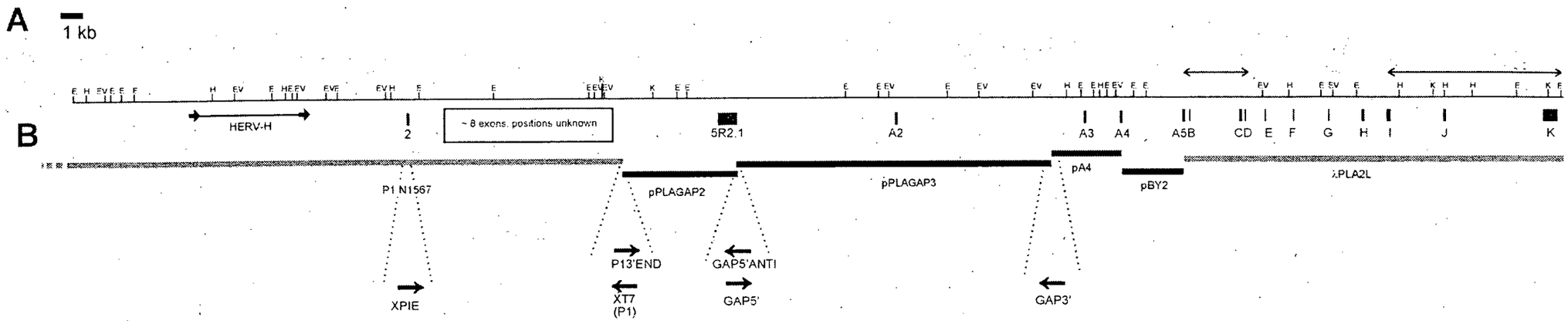


Figure 3.2 Genomic map of the human PLA2L locus.
(**A**). A partial restriction map showing the location of EcoRI (E), EcoRV (EV), KpnI and HindIII (H) sites, respectively. The initial "genomic gap" region, spanning from exon "2" (2 in Fig.3.1B) to exon A2 has not been mapped with HindIII, and the genomic clones downstream of exon "2" have not been mapped with KpnI. Due to complex HERV-H recombination events, not all restriction sites are known within the HERV-H element. The HERV-H retroelement and PLA2L exons are shown below the restriction map. Exons shown in Table 4 are indicated with the corresponding letter below each exon. Regions encoding the PLA-like domains are indicated by double-ended arrows. (**B**). Genomic clones used for mapping and sequencing. P1 bacteriophage clone N1567 and lambda bacteriophage clone λPLA2L are indicated by grey bars. Overlapping long-range genomic PCR-generated plasmid subclones, pPLAGAP2, pPLAGAP3, pA4 and pBY2, are indicated by black bars. Subclone pPLAGAP1 composes the 3' end of the P1N1567 insert and is not shown separately. Primers used to generate the PLAGAP clones are shown as black arrows below dotted lines indicating their position in the genomic clones.

Interestingly, the general intron/exon structure within the PLA-domain regions matches quite closely to that of other $sPLA_2$ genes with small exon size differences, and some larger intron size differences. This is shown in Figure 3.3, which compares the amino acid sequences of the two PLA2L domains to four other $sPLA_2$ for which the genomic structure has been determined. Secreted PLA_2s encode a hydrophobic signal sequence followed by a short propeptide which is cleaved to give the mature enzyme (Seilhamer et al. 1989a). An intron interrupts the signal sequence as shown in Fig. 3.3. Interestingly, although the 5' and 3' PLA2L domains are not similar to other $sPLA_2$s in this region, an intron is located in both domains in a similar region. The 3' boundary of the following exon of the 5' PLA2L (exon A) domain exactly matches both human Group I and II $sPLA_2$ and the splice site of rat Group II and Habu snake venom Group II $sPLA_2$. Exon B is truncated 4 bp on its 3' end relative to the other $sPLA_2$s, and is followed by a large intron similar in size to the approximately 2.6-kb human Group I intron and 2 kb Group II intron. The $sPLA_2$ homology in the 5' domain ends in exon C, with the invariant cysteine at position 127 in Fig. 3.3 being the last conserved residue between Group I $sPLA_2$ and 5' PLA2L, and the reading frame of human Group I ending one residue following the cysteine. Exon C continues for 13 bp, having identity with 2 of 4 amino acids to the slightly longer human and rat Group II enzymes, and then is followed by a 150-bp intron. This small intron is succeeded by an interdomain region of 320 bp of cDNA and approximately 5 kb of genomic DNA (Fig. 3.2). This region contains five very small exons (exons D - H) lacking homology to any known sequence. After the interdomain region, the first exon of the 3' PLA2L domain begins (exon I). This exon contains 49 bases of sequence unrelated to PLA_2 at the 5' end,

Table 4 Exon-Intron Boundaries and Exon and Intron Sizes of PLA-domains of PLA2L

Exon	Position[a]	Exon Size (bp)	5' Splice	Intron Size	3' Splice
					...tctctcccagGAGTCTGTTC
A2[c]	688-734	46	CCCCATGCCGgtgagtaatg	....~9000...	tcacttatagGAGGCCATCC
A3	735-800	65	AACAATATCAgtaagggttc	....~1400...	ttacacttagATATCACTTT
A4	801-857	56	GAAATTTTTGgtaagttaag	...~1900....	accccccacagATTGCCTGGG
A5	858-1032	175	AATCTGACAGgtaagtgagt	....368....	cttcctacagCTGCTGCTTC
B	1033-1145	113	ATCATATGTGGgtaagcatct	..~2600....	ctgtgtgcagAGTCCAAGGA
C	1146-1274	129	CAGACTCCAG[b]gtaggaagaa	..~150....	ttcttttcagAGACAACCAT
D	1275-1316	42	CTGCCCAGAGgtaaggcctc	...~950....	ctttttccagTGGTTCCTGT
E	1317-1367	51	TCAGGAGAAGgtgagccaga	..~1150....	ctgttcacagTGGCTGCAGA
F	1368-1415	48	TCCAAGAAGAgtaacagagg	..~1500....	tggggaacagAAGCAGGCCA
G	1416-1469	54	TCCCCTCCAGgtaagcctac	..~1480....	tcattttcagGATCTGCAGA
H	1470-1595	126	ACTGAAAAAGgtgagcagaa	...~900....	ttcctttcagCCTGTGACAG
I	1596-1767	172	ACCTAGACAGgtattggaga	..~2400....	ttggtcttagGTGCTGCTTG
J	1768-1894	107	ACGCCCAAGTgtaagtgctg	..~4600....	tccttctaagGTGGGGGCCA
K	1875-2413*	>539	CTGTGGAAGC*TCAATAAATACTC		

[a] Nucleotide position relative to PLA2L cDNA clone AF-5, Fig. 2, Feuchter-Murthy et al., 1993.

[b] This sequence supersedes the PLA2L sequence, Fig. 2, Feuchter-Murthy et al., 1993

[c] Exons A2-A5 are additional upstream exons which were subsequently realized to belong to the same transcriptional unit as exons B-K. As exon A2 possesses a splice acceptor site, it cannot be the first exon.

* Corresponds to the 3' end of cDNA clone AF-5.

Underlined sequence corresponds to the putative poly(A) signal for PLA2L

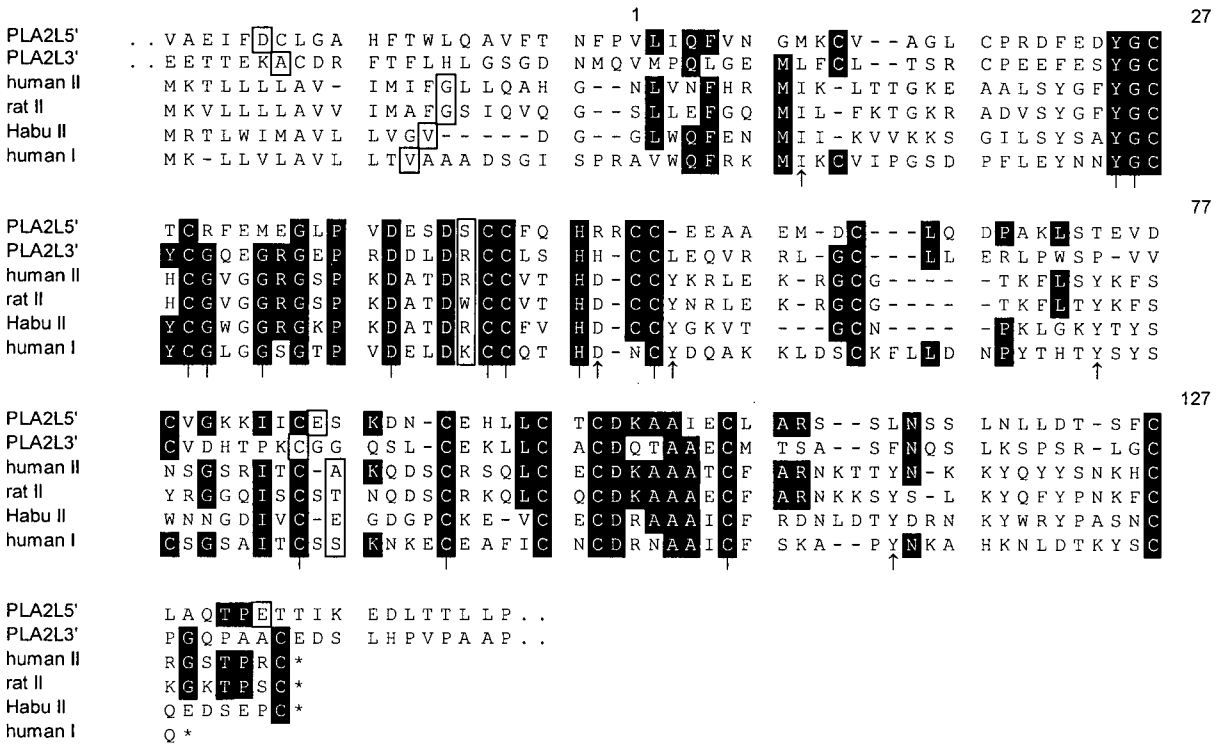


Figure 3.3 Similarity of PLA-domains to sPLA$_2$ and intron/exon structure. Alignment of amino acid sequences of the 5' and 3' PLA2L-domains with other secreted PLA$_2$s. Residues that are conserved in at least 1 PLA2L-domain and at least 2 sPLA$_2$s are shown in reverse print. Residues where intron-exon boundaries occur are shown boxed. The boxed residues in row PLA2L5' correspond to the starts of exons A, B, C, and D, respectively, and those in row PLA2L3' correspond to the start of exons I, J and K, respectively. Exon nomenclature is as in Figure 3.2. Stop codons are shown as asterisks. Arrows indicate 18 conserved positions in all sPLA$_2$s. Numbering is according to the system of (Renetseder et al. 1985), position 1 is the start of both PLA2L-domains and mature, active sPLA$_2$. Dashes indicate gaps introduced for optimal alignment of all six sequences. The start of the PLA2L-domains of PLA2L is indicated with a 1. PLA2L5', 5' PLA-domain; PLA2L3', 3' PLA-domain; Human II, human Group II sPLA$_2$ (Seilhamer et al. 1989a); Rat II, rat Group II sPLA$_2$ (Komada et al. 1990); Habu II, Habu snake venom Group II PLA$_2$ (Nakashima et al. 1993); Human I, human Group I sPLA$_2$ (Seilhamer et al. 1986).

followed by 121 bases of related sequence, splicing at the same point as both human Group I and II $sPLA_2$s (Fig. 3.3) as well as other more diverged Group II $sPLA_2$s. At over 2.4 kb in length, the adjacent intron in the 3' PLA2L region is much larger than both the matching introns from Group I and II PLA_2s, as well as that of the 5' PLA2L domain (Table 4). This large intron is followed by a 107-bp exon J which is 6 bp shorter than its cognate exon in Group I and II PLA_2s, and 3 bp shorter than the exon in the 5' PLA2L domain. The 3' intron is approximately equal in size to other $sPLA_2$s and the corresponding intron in the 5' PLA2L domain, but contains a lengthy CT-repeat region. At 540-bp, the third and last exon of the 3' domain (exon K) is much larger than other PLA2L exons, and continues to the end of cDNA clone AF-5, encompassing the stop codon and the 3' untranslated region. The original 2.4kb PLA2L cDNA, AF-5, was believed to be a slight 3' truncation of a 2.5 kb mRNA detected by Northern analysis, as it lacked both polyadenylation and a poly(A) signal (Feuchter-Murthy et al. 1993). Upon sequencing the genomic clone containing the 3' end of PLA2L, an excellent match to a poly(A)signal was found, just 1 base 3' to the end of the original AF-5 clone (Table 4).

3.2.3 Age of the HERV-H insertion.

As the HERV-H family is primate-restricted (Goodchild et al. 1993; Mager and Freeman 1995), I attempted to determine the approximate time of integration of the element into the PLA2L locus during the course of primate evolution. To do this, I performed genomic PCR on the DNA from a variety of primate cell lines. When a primer set (A1, A2) flanking the HERV-H element was used, a 377 bp product was expected in species lacking the element in the PLA2L gene (Fig. 3.4c). In humans and other species with the HERV-H element within the PLA2L locus, a product of over 6 kb containing the entire element was expected, but as the length is beyond the range for the normal PCR conditions and enzyme used, no product was expected to be seen. On the ethidium bromide-stained gel a product of approximately 370 bp size , was observed in orangutan and a slightly larger product in gibbon, and an unexpected band in human DNA (Fig. 3.4a). Upon hybridization to Probe 4, derived from unique HERV-H flanking sequence, both orangutan and gibbon products hybridized, while no hybridization was seen in the human lane. This indicates that the PCR product amplified from human DNA was non-specific and may be due to the fact that this region is slightly repetitive in the genome (unpublished observations). This result is indicative of the presence of HERV-H within the PLA2L gene in human DNA, and the absence of the element in orangutan and gibbon DNAs, respectively. The deviation from the expected size product in gibbon DNA is explained by a 123-bp length difference previously detected in the region, 3' to the HERV-H insertion site, in both gibbon and baboon but not in orangutan (and therefore unrelated to the HERV-H insertion nearby).

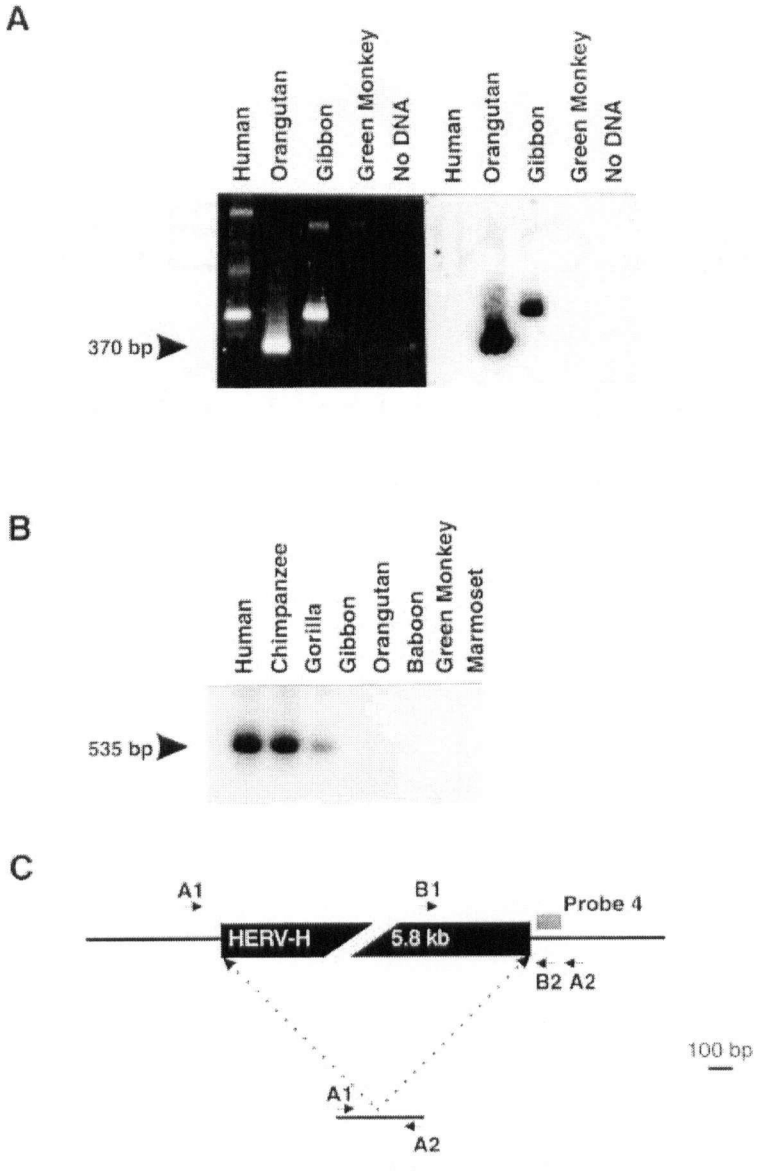


Figure 3.4 Integration time of HERV-H into the PLA2L locus.
Genomic PCR using primers both flanking and within the HERV-H insertion, was performed upon various primate genomic DNAs. Subsequent Southern blots were hybridized to Probe 4. (**A**). Panel 1: the ethidium bromide stained agarose gel showing the results of the PCR using primers A1 and A2 on the indicated genomic DNAs. Panel 2: results of the Southern blotting and hybridization to Probe 4 of the gel in panel 1. (**B**). Results of hybridization of genomic PCR on the indicated primate DNAs, using primers B1 and B2, to Probe 4. A "No DNA" control is not shown in this panel, but was negative. (**C**). Schematic of genomic regions flanking the HERV-H integration, showing primers used. Above is shown the locus in humans, chimpanzees and gorillas, and below is shown the locus in all lower primates.

This region was cloned and sequenced from baboon and showed no homology to any known sequence (data not shown). The absence of an amplification product in African green monkey is likely the result of nucleotide divergence in the primer binding sites rather than a problem with the DNA sample, as all the primate DNA samples used here have been successfully employed in other PCR experiments (Goodchild et al. 1993). To confirm these results, another primate genomic PCR experiment was performed using a primer (B1), derived from the 5' end of a consensus HERV-H LTR (Mager and Freeman 1987), with one derived from unique flanking sequence (B2) (Fig. 3.4c). PCR with these primers is expected to produce a 535-bp product in species possessing the element, with no product being amplified in species lacking the HERV-H element in PLA2L. A wider range of primate DNAs were used in this experiment, and as the PCR used one primer derived from a repetitive element, the gel was blotted and hybridized to a flanking probe since several non-specific products were obtained. Hybridization to Probe 4, directly 3' to the HERV-H element, identified the expected product in human, chimpanzee and gorilla DNAs, respectively, while no product was seen in orangutan, gibbon, and all lower primates (Fig. 3.4b). These data indicate that the HERV-H element present in the PLA2L gene integrated between the divergence of the great ape lineage and orangutan, estimated to be approximately 17 MYA (Sibley and Ahlquist 1987), with orangutan and all lesser primates and mammals lacking the element at the PLA2L locus. Figure 3.5 shows a schematic of the evolutionary radiation of primates, relative to the consequent expansions in HERV-H copy number. HERV-H elements cannot be detected in primate ancestors

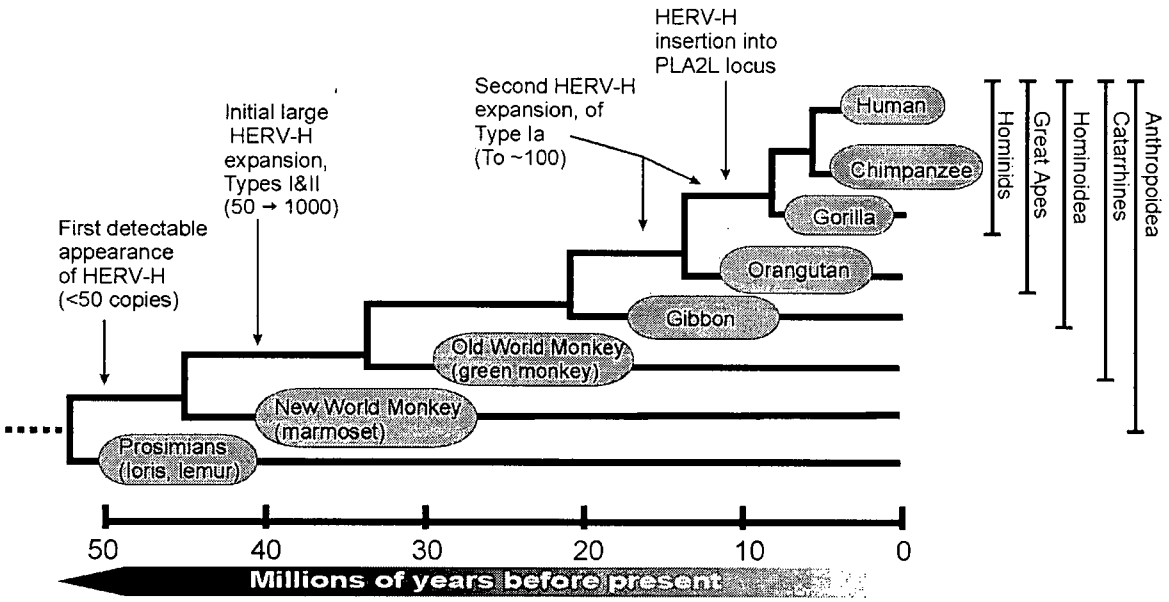


Figure 3.5 Primate speciation and radiation relative to HERV-H expansion. A phylogenetic tree of primate evolution. All branchpoint timings are estimates based on comparative genetics and calibrated using the fossil record (Kelley 1992; Martin 1992). MYA; millions of years ago. The common ancestor to prosimians and simians existed 65 to 56 MYA. Common family and order names are shown at right. The important evolutionary expansions of HERV-H are shown with arrows.

further diverged than New World monkeys (Goodchild et al. 1993), and not other mammals such as dog or mouse.

3.2.4 Chromosomal Localization of PLA2L

Human chromosomal localization of PLA2L was undertaken using PCR on a complete NIGMS rodent/human somatic cell hybrid mapping panel, using a primer set 3' to the HERV-H element (PLAMAP/S42A). An expected product of 585 bp was seen in the chromosome 8 hybrid, the human genomic DNA positive control, and faintly in the chromosome 20 hybrid (Fig. 3.6a). This faint signal was due to the chromosome 20 hybrid containing a small amount of chromosome 8, among others (Drwinga et al. 1993). PCR was then performed on DNAs from a panel of chromosome 8 deletion somatic cell hybrids (Wagner et al. 1991; Wood et al. 1992), which was performed in the laboratory of Dr. S. Wood, Medical Genetics, University of British Columbia. The expected 585 bp product was seen only in hybrids MGV270 and MGV271 (data not shown), whereas all other hybrids were negative. The minimum distance covered by these hybrids encompasses bands 8q24.1 to the terminal 8q24.3, thereby localizing PLA2L to the telomeric end of the long arm of human chromosome 8 (Fig. 3.6b). The chromosome 8q localization of PLA2L differs from the localization of other human $sPLA_2$s, with the Group I gene localized to chromosome 12q (Seilhamer et al. 1989b) and the Group II PLA_2 gene being part of a suspected multigene cluster on human chromosome 1p (Johnson et al. 1990; Tischfield et al. 1996).

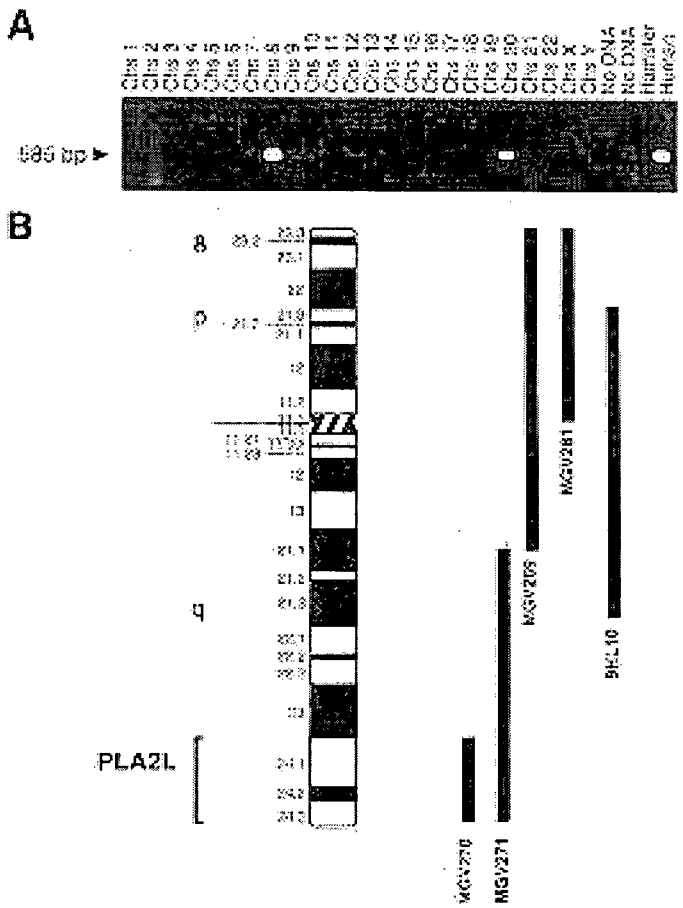


Figure 3.6 Chromosomal and regional localization of PLA2L.
(**A**). Assignment of PLA2L to human chromosome 8. PCR using primer set PLAMAP/S42A was performed upon a panel of human/rodent monochromosomal somatic cell hybrids (Drwinga et al. 1993). Lanes labeled 1-22 are hybrids of the 22 human autosomes. X and Y are hybrids with the X and Y chromosomes and N is a negative control. Ham is total hamster DNA and hum is total human DNA. The expected product of 585-bp was seen in the human chromosome 8 hybrid and the total human genomic DNA control. A faint band visible in the chromosome 20 hybrid was due to the presence of a small amount of chromosome 8 (Drwinga et al. 1993). (**B**). An idiogram of human chromosome 8 showing the regional mapping of PLA2L to 8q24.1 - 8q24.3. PCR using primer set PLAMAP/S42A was performed on the chromosome 8 deletion hybrids shown at the right. The expected band was seen only in MGV270 and MGV271 (Wood et al. 1992).

3.2.5 Evolutionary conservation of PLA2L

To help determine if the PLA2L gene may have an important function in mammals, genomic Southern analysis was performed to detect evidence of sequence conservation in different species. Figure 3.7 shows the results of such an analysis using Probe 3 from the 3' PLA2L domain (this genomic southern was performed by Dixie Mager). Uniquely hybridizing fragments were observed in all mammals tested, including New World monkey (marmoset), prosimian (slow loris), dog and mouse. The two fragments seen in human and other species are due to the fact that the probe spans more than one exon. A similar result was obtained with a probe from the 5' PLA2L domain (data not shown). These results indicate that at least parts of the PLA2L gene have been conserved in mammalian evolution.

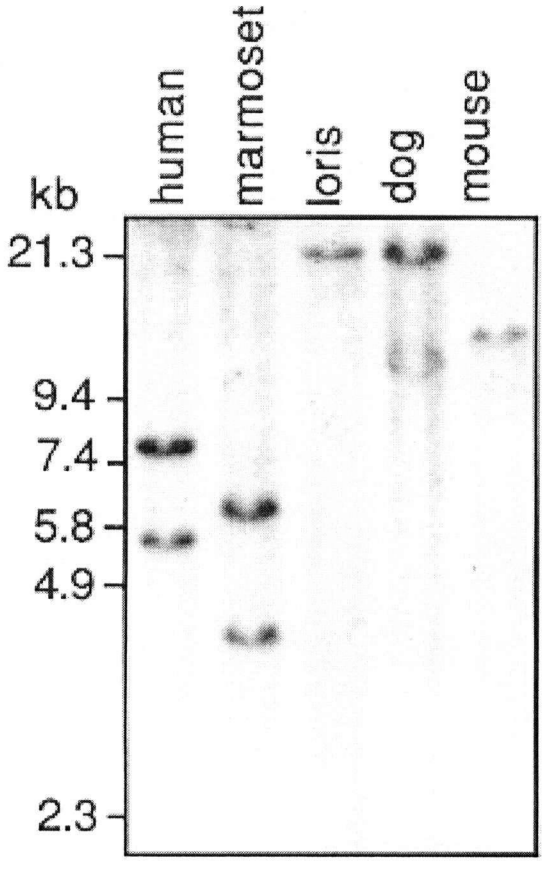


Figure 3.7 DNA sequence conservation of PLA2L.
Genomic DNAs from human, marmoset, slow loris, dog and mouse were digested with EcoRI, electrophoresed and Southern blotted. The filter was hybridized to Probe 3 (Fig. 3.1).

3.3 Discussion

In this study, I have determined the genomic organization and chromosomal localization of the human PLA2L locus and revealed that the basic intron/exon structure of the PLA-domains parallels the structure of active $sPLA_2$ genes. In addition, the PLA2L gene was shown to reside on human chromosome 8. I also examined the evolutionary age of the HERV-H element present within the PLA2L gene, and demonstrated that HERV-H integration at this locus occurred approximately 17 million years ago. Although these experiments do not address PLA2L expression or function, my hypothesis is that the retroviral element has assumed transcriptional control of this gene in humans and great apes, within cells in which the LTR promoter is active.

Beside the amino acid similarity between the PLA-domains and $sPLA_2$, I have shown here that the intron/exon structure of both domains is strikingly similar to the structure of secreted Group I and II PLA_2s. The first exon in both PLA-related domains originates upstream of the start of PLA_2 homology (Fig. 3.3). Interestingly, although the first exon of known $sPLA_2$ genes are not similar to those of the PLA-domains in PLA2L, the relative position of the first intron is conserved. In the case of $sPLA_2$ genes, this intron interrupts the leader sequence which is proteolytically cleaved to produce the active form of the enzyme. Despite the lack of sequence similarity in the leader region, the conserved positioning of introns supports the hypothesis that PLA2L is a precursor of contemporary $sPLA_2$s or that both were derived from another ancestral form.

This hypothesis is also supported by the fact that both PLA-domains share characteristic features of both Group I and II $sPLA_2$s, and therefore cannot be placed

unambiguously in either group (Feuchter-Murthy et al. 1993). Since both domains are highly diverged from each other (37% protein identity, 50% DNA identity), the duplication event giving rise to both PLA-domains must have necessarily been ancient (Feuchter-Murthy et al. 1993). Dennis and co-workers (Davidson and Dennis 1990) have proposed an evolutionary scheme for $sPLA_2$, subsequently updated by Tischfield to reflect newly discovered $sPLA_2$ groups (Tischfield 1997). This scheme entails a series of duplication events starting with a progenitor $sPLA_2$, which undergoes gene duplication to produce a Group I/II precursor and a Group III (bee venom $sPLA_2$) precursor. The Group I/II precursor, thought to be the direct progenitor to Group V due to the maintenance of 12 Cys residues, is again duplicated to generate proper Group I and II enzymes, both possessing 14 Cys residues. The Group II precursor is again duplicated, to produce a 16 Cys-containing Group IIC PLA_2 gene (Tischfield 1997). As the PLA2L PLA-domains have 16 and 17 Cys residues, respectively, it is likely that PLA2L is more closely related to Group IIC genes. As both reptiles and mammals possess various $sPLA_2$s of both groups, these duplications all occurred before their divergence. Dendrograms of amino acid comparisons between both groups of $sPLA_2$ and the PLA-domains of PLA2L show that both domains are as different from all other $sPLA_2$ as Group I is from Group II (Feuchter-Murthy et al. 1993). This implies that the PLA-domains within PLA2L are at least as old as the Group I/II precursor, and therefore likely arose before mammalian/reptile divergence. The extreme age of the domains is supported by the chromosomal localization of PLA2L. PLA2L has been shown to reside upon human chromosome 8, rather than chromosome 1 in the case of Group II $sPLA_2$ and chromosome 12, in the case of Group I. This supports the hypothesis of an ancient duplication event, followed by further duplications and

translocations of the domains, to result in independent evolution of $sPLA_2$s on different chromosomes. Intriguingly, a recently cloned novel $sPLA_2$ has been shown not to unequivocally belong to Groups I or II, and to possess some similar features to the PLA2L PLA-domains, namely the number of Cys residues (16) and the position of the resulting disulfide bridges (Cupillard et al. 1997). Although studies on Group X $sPLA_2$ are in their infancy, it seems probable that a gene in this group is the most closely related of any known gene, to the ancestral progenitor gene whose duplication gave rise to the PLA-domains of PLA2L.

The absolute conservation, in both PLA-domains, of all cysteine residues required for correct $sPLA_2$ secondary structure (Fig. 3.3) implies some form of positive selection for maintenance of the three-dimensional $sPLA_2$ structure. Although these structures would exist *within* the larger PLA2L protein which likely has structure of its own, it is possible that proteolytic cleavage could liberate the PLA-domains. Despite the presumed ancient origin of PLA2L and its unusual structure, the conservation of the reading frame, as well as the positioning of cysteine residues, implies selection upon PLA2L function. In addition, I have shown sequence conservation in other mammals (Fig. 3.7). In the great apes, the HERV-H retroviral integration may have resulted in the abrogation of the function of PLA2L. The pressure to maintain the ORF would have ceased at that time. It can then be calculated, based upon nucleotide substitution rates for non-coding DNA (Li and Graur 1991), that on average, 4 nucleotide substitutions causing termination codons in PLA2L's 1.7kb open reading frame should have occurred. The maintenance of the ORF therefore suggests that the human gene still has a function.

Although phospholipase A_2 enzymatic catalysis can occur in other PLA_2s without the consensus $sPLA_2$ catalytic site (Zupan et al. 1992), it has not been shown to occur in this manner in enzymes sharing homology with $sPLA_2$. It is expected that PLA2L will lack PLA_2 catalytic activity, due to amino acid substitutions in positions required for active site formation and Ca^{2+}-binding. Of the 18 absolutely conserved residues in all active $sPLA_2$s (Davidson and Dennis 1990), both PLA-domains are substituted at 7 of them (Fig. 3.3). The most notable substitution is the highly conserved acidic Asp^{49} which binds Ca^{2+} in the $sPLA_2$ active site (Li et al. 1994), to the basic His^{49}/Arg^{49} found in the respective 5' and 3' PLA2L domains (Fig. 3.3). Interestingly, non-Asp^{49} $sPLA_2$ homologues exist in snake venom, where, although enzymatically inactive, they serve as inhibitors or chaperones for normal $sPLA_2$ (Maraganore and Heinrikson 1986; Davidson and Dennis 1990). This suggests PLA2L could function synergistically as a modulator of endogenous $sPLA_2$, perhaps as a protective protein preventing proteolytic degradation or inhibitor binding.

Endogenous retroviral elements and other retroelements are known to be causative agents of genetic change (Wilkinson et al. 1994; Di Cristofano et al. 1995; Schulte et al. 1996). This change is often mediated through the integration of retroviral elements into cellular genes. For example, the retroviral integration into an intron of the murine *dilute* gene has been shown to alter *dilute* expression, in a tissue specific manner (Seperack et al. 1995). If integrated into exon sequences, ERVs can cause premature truncation by introducing in-frame premature stop codons, or lead to gene: ERV fusions. ERVs integrated within introns can influence adjacent genes by causing aberrant promotion or enhancement, premature polyadenylation, and can, via splicing, cause fusion transcripts with LTR or internal sequences. It is likely that the insertion of

a heterologous HERV-H promoter into a PLA2L intron caused changes in gene expression, but perhaps in a very limited cell type. As HERV-H LTR promoters are most active in primitive cells such as teratocarcinomas, it is presumed that any perturbations in PLA2L gene expression would be most pronounced in, or perhaps exclusive to, these cell types, such as NTera2D1, from which the PLA2L cDNA was isolated. Given the general lack of PLA2L transcription in all cells examined except teratocarcinomas (and including the similar lack of ESTs in all databases), the possibility does exist that the PLA2L fusion transcript is, by nature of its HERV-H LTR promoter, teratocarcinoma-specific.

CHAPTER FOUR: CLONING AND CHARACTERIZATION OF THE MURINE HOMOLOGUE OF PLA2L, OTOCONIN-90, AND DEVELOPMENT OF A HYPOTHESIS FOR PLA2L BIOGENESIS

Sections of this chapter have been drafted into a manuscript:

Wang, Y., Kowalski, P. E., Thalmann, I., Ornitz, D. M., Mager, D. L., and Thalmann, R. Mammalian otoconin-90 encodes a phospholipase A_2 homologue. Submitted.

4.1 Introduction

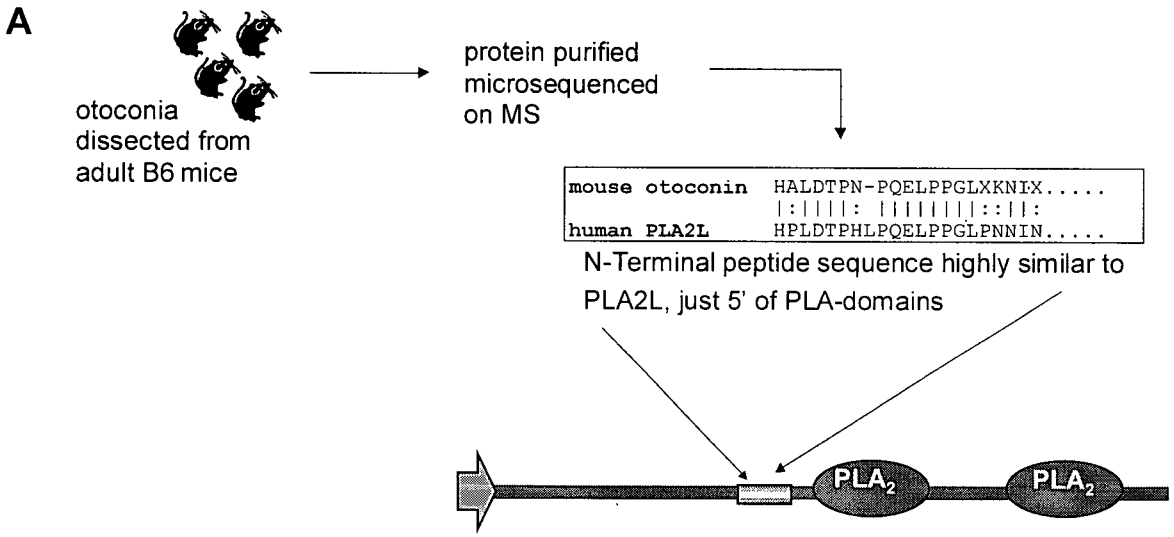
Comparative genomics is one of the most powerful tools biologists possess to examine gene structure, with millions of years of evolution functioning as a powerful screen to identify important domains, sequences and other structural features. As a consequence of the experiments detailed in Chapter 3, the evolutionary time of HERV-H integration into PLA2L was determined to be approximately 17 MYA, predicting that all mammals more diverged than, and including, orangutan, would lack the HERV-H element in the PLA2L genomic locus (Fig. 3.4). The study of the PLA2L transcript under control of a presumably native, endogenous promoter (rather than an exogenous LTR promoter) is the necessary addendum to determine the effects that HERV-H has had upon the expression of the adjacent gene, PLA2L. It follows that the need for comparative genomics would have been obviated had HERV-H-independent PLA2L expression in a human cell line or tissue been found, which was not, after exhaustive searching of greater than 50 tissues. Southern blotting of DNA from evolutionarily divergent species ("zoo-blot") using a fragment of human PLA2L cDNA revealed conservation in evolution as far distant as mouse (Fig. 3.7), or 60 MYA. Conservation of DNA sequence between human and mouse at a level detectable by DNA:DNA hybridization (approximately 70%) usually suggests a necessary function. These results, taken together, led me to attempt to clone the murine homolog of PLA2L by hybridization to human probes at reduced stringency, with the aim of comparing and contrasting the expression of PLA2L from native and HERV-H promoters.

Concurrently to these initial murine genomic cloning experiments, a fortuitous discovery was made by Dr. David Ornitz, Washington University School of Medicine,

regarding the murine homolog of PLA2L. In an effort to identify the primary protein component of murine otoconia, an extracellular bioorganic crystalline structure found in the vestibular system of the inner ear, the most abundant protein was purified and N-terminally sequenced, and found to have distinct similarity (72% identity) to PLA2L (Fig.4.1). This was the first evidence of expression outside of the original human teratocarcinoma cell lines. Surprisingly, the region of homology was found not to lie at the predicted N-terminus of the human PLA2L protein, but in the middle of the predicted protein, in a unique region just upstream of the first PLA-domain (Fig. 4.1a). Otoconia are small dense crystals, composed of calcium carbonate and one major and up to 3 minor glycoprotein components (Pote and Ross 1991). Otoconia are found in the utriclular and saccular macula regions of the inner ear, where they are embedded in the otoconial membrane. The function of the sensory macula is to transduce not sound, but spatial orientation relative to gravity. The function of otoconia in the mammalian inner ear is to add mass to the otoconial membrane, enhancing the sensory perception and sensitivity of the underlying sensory hair cells (Friedmann 1976; Pote and Ross 1991). Mutations affecting otoconia biosynthesis result in loss of equilibria and balance, and circling behavior in mutant mice strains (Lyon et al. 1996). As otoconin is secreted, the N-terminus of the otoconin-90 protein lacks an initiating methionine and presumed secretion signal sequence.

In this chapter I report the genomic and cDNA cloning of otoconin-90, the murine homologue of the downstream transcriptional unit comprising PLA2L, and compare the human and mouse sequences. In addition, I present data supporting the hypothesis that PLA2L is a HERV-H-induced tripartite fusion transcript composed of HERV-H sequences, approximately 8-10 exons of an unknown but conserved gene of

very limited expression, termed HHAG-1, and the adjacent human otoconin-90 ortholog. The current hypothesis entails both adjacent genes being independently expressed (the lack of human EST representation infers that these genes are expressed in a very restricted spatial/temporal pattern) and only being fused in teratocarcinoma lines to generate the PLA2L transcript. As HERV-H LTR promoters are highly active in these cell lines, I hypothesize the synergistic effects of this active heterologous LTR promoter combined with transcriptional antitermination and subsequent intergenic splicing (Fears et al. 1996) as a mechanism of PLA2L biogenesis.



B

```
TACAGGGGCCCATGCTCTGGACACACCAAATCCCCAGGAATTGCCTCCAGGACTGTCAAAAAATATAAGTAAGA
---------+---------+---------+---------+---------+---------+---------+----BH400
ATGTCCCCGGGTACGAGACCTGTGTGGTTTAGGGGTCCTTAACGGAGGTCCTGACAGTTTTTTATATTCATTCT

....G A H A L D T P N P Q E L P P G L S K N I N......
    | | | | | | | | | | | | | | | | | | |
    G A H A L D T P N P Q E L P P G L S K N I x      peptide
```

Figure 4.1 Schematic of murine otoconin-90 discovery and relation to PLA2L. **(A)** Otoconia were dissected from the inner ears of adult mice, protein extracted in EDTA, and analyzed by 2D SDS-PAGE. Otoconin-90 was seen to compose >90% of total otoconial protein, and was directly N-terminally sequenced. The protein sequence was used to search the Genbank databases with the BLAST program, with the only significant match being to human PLA2L, in the middle of the predicted protein, just N-terminal to the first PLA-domain. Performed by Y. Wang, Washington University. **(B)** The relevant exon-containing section of the BH400 genomic subclone sequence showing the its conceptual translation (above) and perfect match to the purified murine otoconin-90 peptide (below). The final residue in the peptide was unsequencable and shown as an X. Examination of the coding exon shows an asparagine (N) residue in that position, in the context of the consensus Asn-X-Thr ,which therefore likely possesses N-linked glycosylation. Glycosylation is known to impair protein sequencing.

4.2 Results

4.2.1 Identification and genomic cloning of murine homologue

Evidence of PLA2L DNA sequence conservation in mouse was first obtained using cross-species genomic Southern blotting, as shown in Figure 3.7. Notably, this Southern shows that the homologous gene in mouse appears to exist as a single copy. The peptide sequence which was 72% identical to PLA2L (nt 745-897) was obtained from N-terminal microsequencing of homogenous murine otoconin-90 (Wang et al. submitted). Using ^{32}P-labeled human cDNA probes, Probes 2 and 3 (Fig. 3.1), a gridded RLDB C57B6 mouse genomic P1 library (Zehetner and Lehrach 1994) was screened at low stringency, and two positive clones were obtained. In addition, a 129 mouse genomic λ library was screened and two clones positive for Probe 3 were plaque purified. As the P1 clones contained far larger inserts (70-100 kb), the λ clones were not characterized further. Purified P1 DNA was probed, again at reduced stringencies, i.e. 55°C, with successive probes along the length of the PLA2L transcript. It was determined that the 5'-most human probe that either P1 clone hybridized to was Probe 5, that both clones were positive for Probe 2 and that while both clones hybridized to Probe 3, clone P1219 was positive for a ~13kb EcoRI fragment and P1200 was positive for a ~9 kb EcoRI fragment. Clone P1219 was used in all further experiments, as restriction mapping and hybridizations determined that it contained a larger portion, and perhaps all, of the otoconin-90 genomic locus. As confirmation was needed that the genomic clones actually contained the gene coding for otoconin-90, the human probe containing the homologous region to the peptide sequence, Probe 5 (Fig. 3.1) was used to screen subclones of HindIII digested P1219 DNA. A 5.8 kb HindIII fragment, positive for Probe 5, was isolated and termed 308H. A

number of restriction enzyme double digests were performed on this clone, and the smallest fragment seen to hybridize to Probe 5 was a 400 bp BamHI/HincII product. This fragment, BH400, was subcloned and sequenced, and found to contain an exon coding for the original purified peptide. This is shown in Figure 4.1b, with the relevant genomic sequence at the top, its conceptual translation and alignment to the otoconin-90 peptide below. This result served as confirmation that the P1 clone contains the genomic locus of the otoconin-90 gene, the ortholog of PLA2L.

4.2.2 Construction of an otoconin-90 murine genomic contig

To further characterize the genomic locus, an approach of isolating overlapping subclones from the large-insert P1219 to generate a contig was used. Overlapping subclones were generated from P1219 DNA by using fragments from ends of existing clones to probe "mini-libraries" of randomly cloned P1 restriction fragments, or by long-range PCR on P1219 DNA using primers derived from otoconin-90 cDNA sequences. In this manner, the subcloned genomic locus was extended 6 kb upstream from the original 308H clone, and downstream 28 kb.

4.2.3 cDNA cloning and consensus sequence assembly

As of July 1998, approximately three-quarters of the otoconin-90 transcript was represented as Merck/Washington University EST clones W50767, AA034721 and AA437511. All three clones were derived from E13-E14 mouse embryo cDNA libraries, which was the first window into the developmental regulation of this gene; otoconin-90 was subsequently shown by Y. Wang to be maximally expressed between E14.5-E17 embryonic stages. The experimental steps which led up to the generation

of a nearly full length cDNA sequence for otoconin-90 are shown in Figure 4.2. The downstream two-thirds were primarily derived from EST clones, while the upstream third was determined by cDNA cloning. The genomic exon (likely the 2nd) found in the BH400 clone served as the "anchor" for further cloning. The gap between this exon and the 5' end of the EST AA034721 was cloned by performing RT-PCR upon E14 mouse embryo head total RNA, using primers from the 3' end of the 2nd exon and the 5' end of the EST AA034721 sequence, as detailed in Chapter 2.

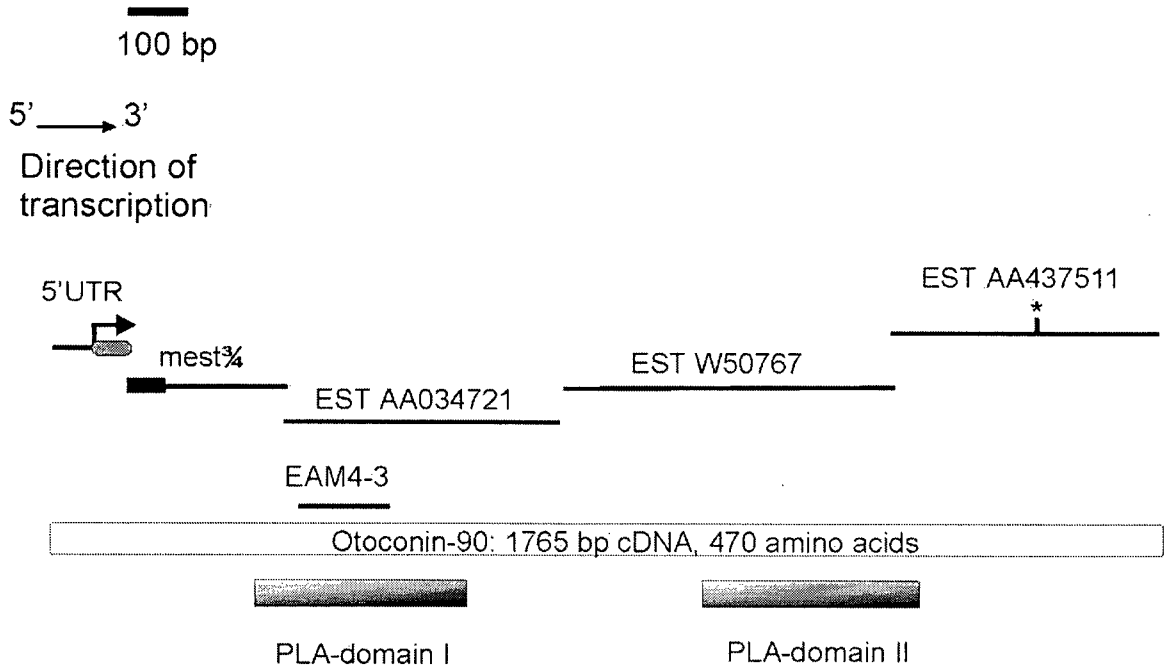


Figure 4.2 Assembly of murine otoconin-90 cDNA contig.
The seven cDNA clones making up the consensus otoconin-90 cDNA sequence are shown as black bars. The three EST clones include publicly available Genbank sequence reads in addition to manual sequencing of gaps or poor sequence. The start codon is shown as a bent arrow, while the stop codon is shown as an asterisk. The predicted signal sequence is drawn as a gray oval immediately following the start codon. The original otoconin-90 peptide is shown as a black rectangle; the initial BH400 genomic clone is contained within this sequence. The clone 5' UTR was generated by nested PCR on a mouse E14 cDNA library, and is described in section 4.2.4. The mest ¾ clone was created using RT-PCR upon E14.5 mouse embryo head RNA as detailed in Chapter 2, using primers from the initial BH400 exon and from the 5' end of the EST AA034721 sequence. EAM4-3 is a exon clone generated prior to the sequencing of EST AA034721 using the exon amplification procedure upon P1219 genomic DNA.

```
   1  CGATCATTCAACAGCCTTCGAGTGGAGCTTCACTTCGCGGAAGCTGCTCTAGCCTATGCC   60
  61  TACACCTTGTCCTCTGCACTGCCACAATGATTATGCTGCTCATGGTCGGTATGCTGATGG   120
                                  M  I  M  L  L  M  V  G  M  L  M  A
 121  CCCCCTGTGTTGGGGCCCATGCTCTGGACACACCAAATCCCCAGGAATTGCCTCCAGGAC   180
       P  C  V  G  A  H  A  L  D  T  P  N  P  Q  E  L  P  P  G  L
 181  TGTCAAAAAATATAAATATCACTTTCTTCAATGGAGTGTTTAAAAACGTGGAAAGTGTGG   240
       S  K  N  I  N  I  T  F  F  N  G  V  F  K  N  V  E  S  V  A
 241  CTGAAATTTTTGACTGCCTAGGTTCCCACTTCACCTGGCTGCAGGCTGTCTTCACCAACT   300
       E  I  F  D  C  L  G  S  H  F  T  W  L  Q  A  V  F  T  N  F
 301  TCCCGCTGCTCCTCCAGTTTGTAAACAGTATGAGGTGTGTAACTGGCCTCTGCCCCCGGG   360
       P  L  L  L  Q  F  V  N  S  M  R  C  V  T  G  L  C  P  R  D
 361  ACTTTGAAGACTATGGTTGTGCCTGTAGGTTTGAGATGGAAGGGATGCCTGTGGACGAGT   420
       F  E  D  Y  G  C  A  C  R  F  E  M  E  G  M  P  V  D  E  S
 421  CTGATATCTGCTGCTTCCAGCATCGCAGGTGCTATGAGGAAGCTGTTGAGATGGACTGTC   480
       D  I  C  C  F  Q  H  R  R  C  Y  E  E  A  V  E  M  D  C  L
 481  TCCAAGACCCTGCCAAGCTCAGTGCAGATGTGGATTGCACCAACAAACAGATCACATGTG   540
       Q  D  P  A  K  L  S  A  D  V  D  C  T  N  K  Q  I  T  C  E
 541  AGTCCGAGGATCCCTGTGAGCGTCTACTGTGTACGTGTGACAAGGCTGCTGTGGAGTGCC   600
       S  E  D  P  C  E  R  L  L  C  T  C  D  K  A  A  V  E  C  L
 601  TGGCTCAGTCTGGCATCAACTCCTCCCTGAACTTCCTGGATGCTTCCTTCTGCCTCCCTC   660
       A  Q  S  G  I  N  S  S  L  N  F  L  D  A  S  F  C  L  P  Q
 661  AGACTCCAGAGACAACTAGTGGGAAGGCTGCAACACTGTTGCCTAGAGGGATTCCTGAAA   720
       T  P  E  T  T  S  G  K  A  A  T  L  L  P  R  G  I  P  E  K
 721  AGCCCACAGATACCAGTCAGATAGCCCTGTCAGGAGAAGAGTCTGTTCAAGATCTTCAAG   780
       P  T  D  T  S  Q  I  A  L  S  G  E  E  S  V  Q  D  L  Q  D
 781  ACACACAAGCTTCTAGGACCACATCAAGTCCAGGATCTGCAGAGATTATTGCCCTAGCCA   840
       T  Q  A  S  R  T  T  S  S  P  G  S  A  E  I  I  A  L  A  K
 841  AAGGTACAACCCACTCTGCTGGCATCAAACCACTGAGGTTGGGAGTCTCATCTGTTGACA   900
       G  T  T  H  S  A  G  I  K  P  L  R  L  G  V  S  S  V  D  N
 901  ATGGTTCCCAGGAAGCAGCTGGAAAAGCAGCCTGTGACAGATTGGCCTTCGTGCATCTGG   960
       G  S  Q  E  A  A  G  K  A  A  C  D  R  L  A  F  V  H  L  G
 961  GTGATGGGGACAGCATGACGGCCATGCTGCAGCTTGGAGAGATGCTCTTCTGTCTAACAT   1020
       D  G  D  S  M  T  A  M  L  Q  L  G  E  M  L  F  C  L  T  S
1021  CCCATTGCCCAGAGGAATTTGAAACTTACGGCTGCTACTGTGGAAGAGAAGGAAGAGGAG   1080
       H  C  P  E  E  F  E  T  Y  G  C  Y  C  G  R  E  G  R  G  E
1081  AGCCAAGGGACACCCTGGATAGGTGCTGTCTGTCCCATCACTGCTGTTTGGAGCAGATGA   1140
       P  R  D  T  L  V  R  C  C  L  S  H  H  C  C  L  E  Q  M  R
1141  GACAAGTGGGCTGCCTCCATGGAAGGCGTTCTCAGTCATCTGTGGTATGTGAAGACCACA   1200
       Q  V  G  C  L  H  G  R  R  S  Q  S  S  V  V  C  E  D  H  M
1201  TGGCCAAATGTGTGGGGCAGAGCCTGTGTGAGAAGCTACTATGTGCCTGTGACCAGATGG   1260
       A  K  C  V  G  Q  S  L  C  E  K  L  L  C  A  C  D  Q  M  A
1261  CAGCTGAGTGCATGGCCTCTGCCTTCTTTAATCAAAGCCTCAAGTCACCAGACGGAGCCG   1320
       A  E  C  M  A  S  A  F  F  N  Q  S  L  K  S  P  D  G  A  E
1321  AGTGTCAAGGCGAGCCTGTGTCCTGTGAGGATGGCATGCTCCAGGGCACCTTGGCCTCTT   1380
       C  Q  G  E  P  V  S  C  E  D  G  M  L  Q  G  T  L  A  S  S
1381  CTGTGGACTCCAGTTCTGAGGAGAATAGTGAGGAAGCTCCACCGCAGATGGAACGCCTAA   1440
       V  D  S  S  S  E  E  N  S  E  E  A  P  P  Q  M  E  R  L  R
1441  GAAGATTTCTGGAAAAGCCTCCTGGTCCCTTGGGGGCCAGACCCCTCGGTGGGAAATAAG   1500
       R  F  L  E  K  P  P  G  P  L  G  A  R  P  L  G  G  K  *
1501  ATGCTACGTGCTAGTAGCTCTAAGCTGTCTGAGCCCTTTGGCCCTCAGTCCCACCCATAG   1560
1561  GAGCCTTAGCAGGGTCTCCAAGGGAGCAGGGACAGCCACCCCTTTATCCATGAGTCTCCC   1620
1621  CTTTATCCACGAGCCTCCTGAAACTTGTCAGCACAGATACATGTGTCTGGAGAATAACTG   1680
1681  CAGATGACAGCCCTTTTCTTCCTGTAGTTCACTATGGAAGCTCAATAAATTCTCTATGCC   1740
1741  ACATAAAAAAAAAAAAAAAAAAAAA  1765
```

Figure 4.3 Otoconin-90 cDNA sequence.
Consensus sequence derived from cDNA and EST sequencing. The 5' end of the sequence has been derived from cDNA library PCR, and although no other larger products were seen, it must be noted that this may not represent the true 5' end of the transcript. The predicted signal sequence is shown boxed. The original otoconin-90 peptide sequence is shown in bold, as is the poly(A) signal near the 3' end. The PLA-domains are underlined. As otoconin-90 is known to be glycosylated, potential N-linked glycosylation sites are shown in bold italics. This sequence will be submitted to GenBank upon publication of the associated manuscript.

The cloning and characterization of the 5' UTR clone, and the discovery that it contains, as indicative of expression from a native promoter, an alternative 5' end as compared to PLA2L, is detailed in the following section.

The complete cDNA sequence for murine otoconin-90 encompasses 1765 nucleotides, possessing an open reading frame of 470 residues, and is shown in Figure 4.3. At 86 bp, the 5' UTR is slightly shorter than average length (Kozak 1996), while the 3' UTR, which contains the consensus AATAAA polyadenylation signal at its 3' end, is made up of 246 bases, not including the poly(A) tail. This matches very well in length to the 245 bp 3' UTR of PLA2L, although being quite dissimilar in nucleotide sequence. At the nucleotide level, it is 70% identical (Fig. 4.4), while at the protein level, otoconin-90 is 80% similar and 69% identical to the analogous region of PLA2L (Fig. 4.5). Emphasizing the extreme age and divergence of the PLA-domains compared to contemporary $sPLA_2$ enzymes, each PLA-domain in mouse otoconin-90 is only 32-34% identical to rat $sPLA_2$ protein, and both domains are only 36% / 51% identical to each other at amino acid and nucleotide levels, respectively. Notably, this level of inter-domain identity is exactly concordant to that of the PLA-domains in PLA2L (Feuchter-Murthy et al. 1993). Nucleotide identity between PLA2L and otoconin-90 is largely limited to the coding regions; the 5' and 3' UTRs (and signal sequence to a lesser extent; Fig. 4.9) are quite dissimilar (Fig. 4.4). Pairwise comparisons of the PLA-domains in all known otoconin genes are shown aligned to a representative $sPLA_2$, human Group II (Seilhamer et al. 1989a) in Figure 4.6. The only other otoconin known is derived from the African clawed frog *Xenopus laevis*. *Xenopus* otoconin-22 is much smaller and contains only one PLA-domain, implying the gene duplication giving rise to 2 PLA-domains found in

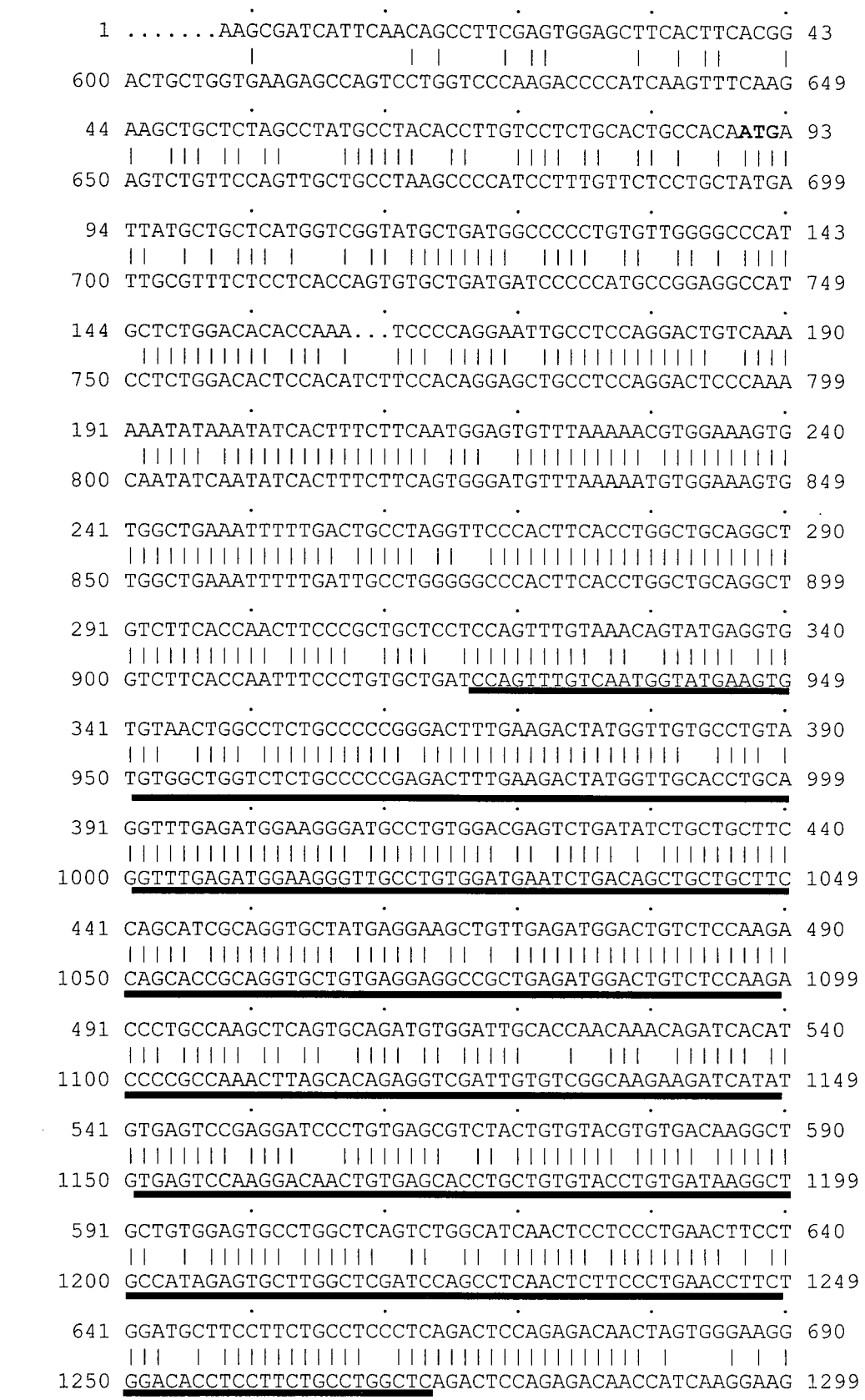


Fig. 4.4 (page 1 of 3)

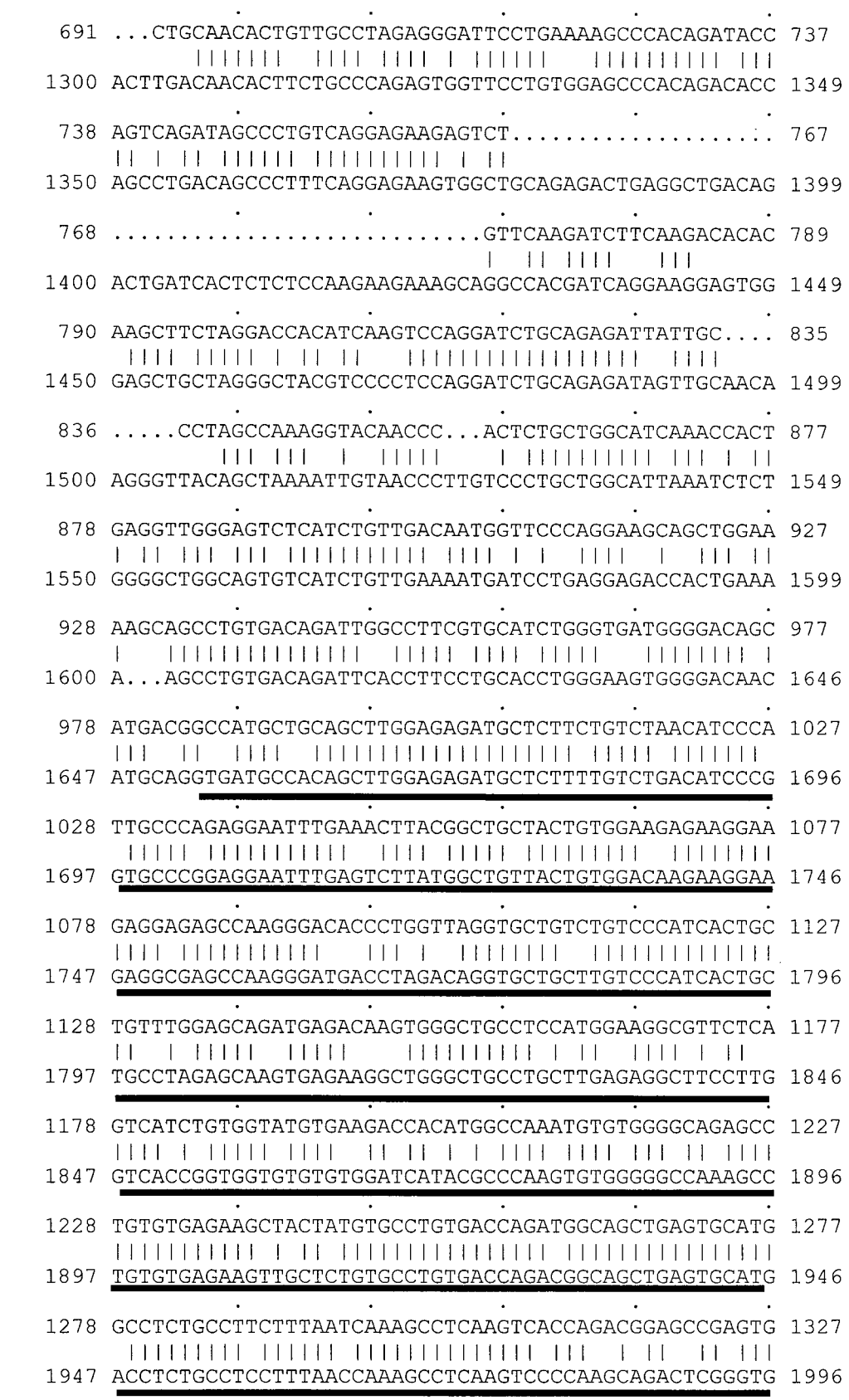


Fig. 4.4 (page 2 of 3)

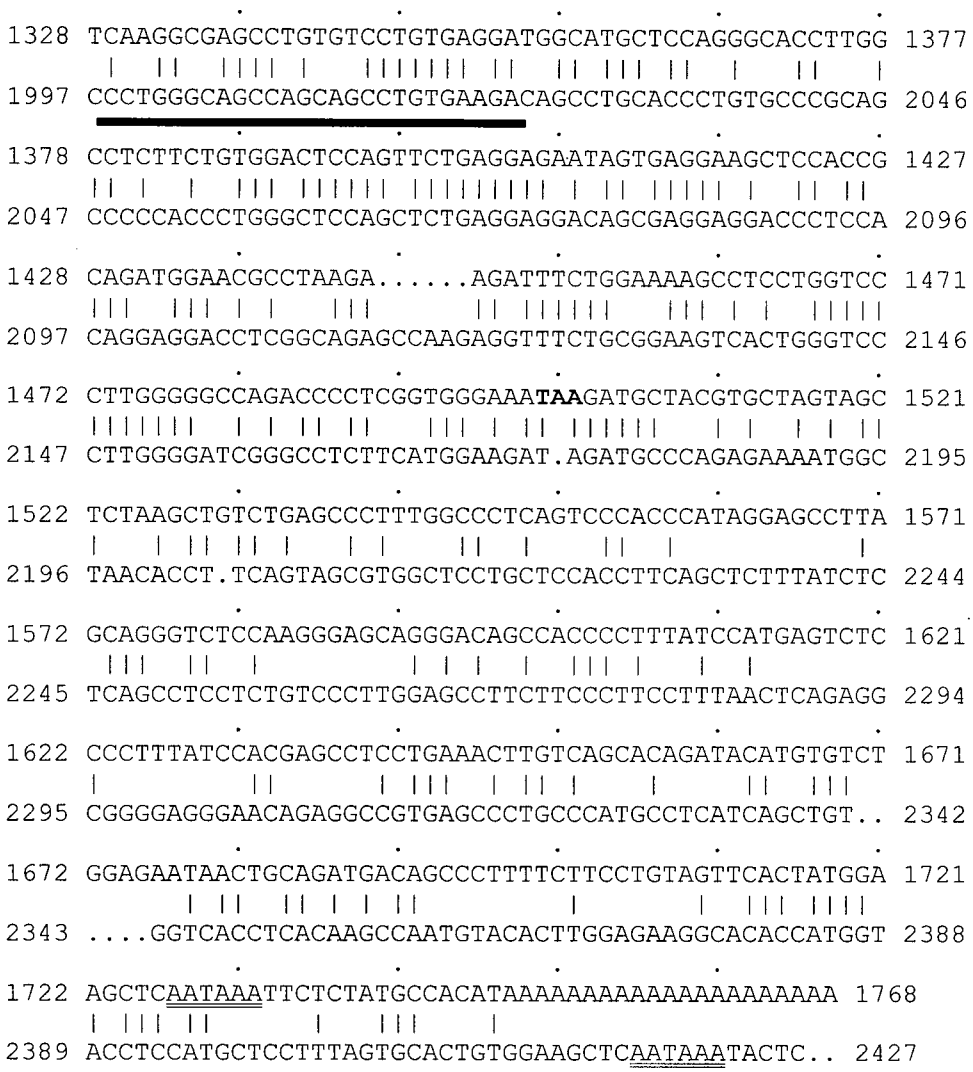


Figure 4.4 Alignment of murine and human otoconin-90 DNA sequence
Alignment of the relevant section of the human PLA2L cDNA to the murine otoconin-90 consensus cDNA, using the GCG GAP program. 70% identity is seen. PLA-domains are underlined. It is notable that the homology appears restricted to the coding regions, with the 3' UTR and to a lesser extent, the 5'UTR being largely nonhomologous. Identity can also be seen to be slightly reduced in the intradomain region between the PLA-domains. Start and stop codons are shown in bold type. Putative polyadenylation signals are double underlined.

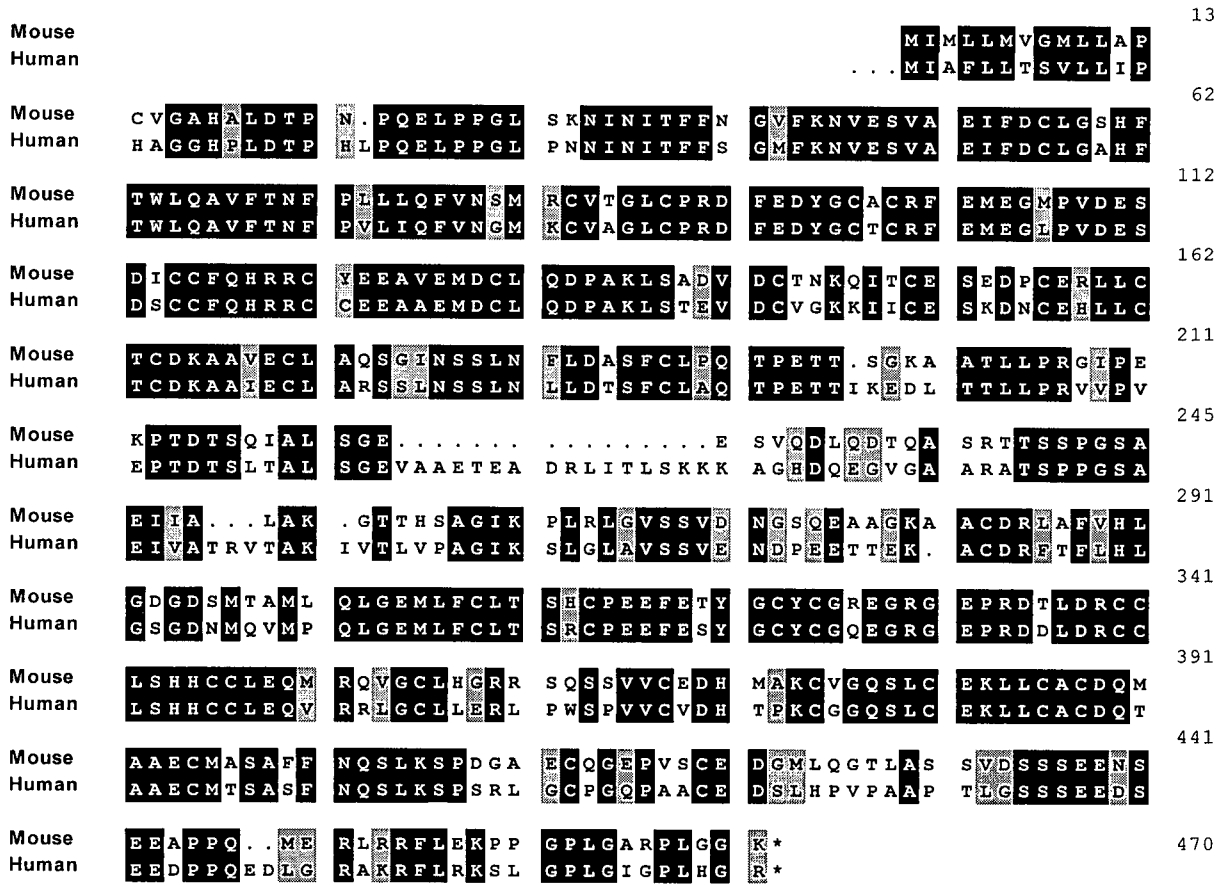


Figure 4.5 Amino acid alignment of murine otoconin-90 and human PLA2L
Pairwise alignment of the complete predicted amino acid sequence of murine otoconin-90 and human PLA2L, using the GCG program GAP. Identity is shown by reverse shading, while conservative substitutions are shown in grey shading. Numbering is relative to the otoconin-90 sequence. Gaps introduced for optimal alignment are shown by dots. Stop codons are shown as asterisks. 80.2% similarity, 69.5% identity.

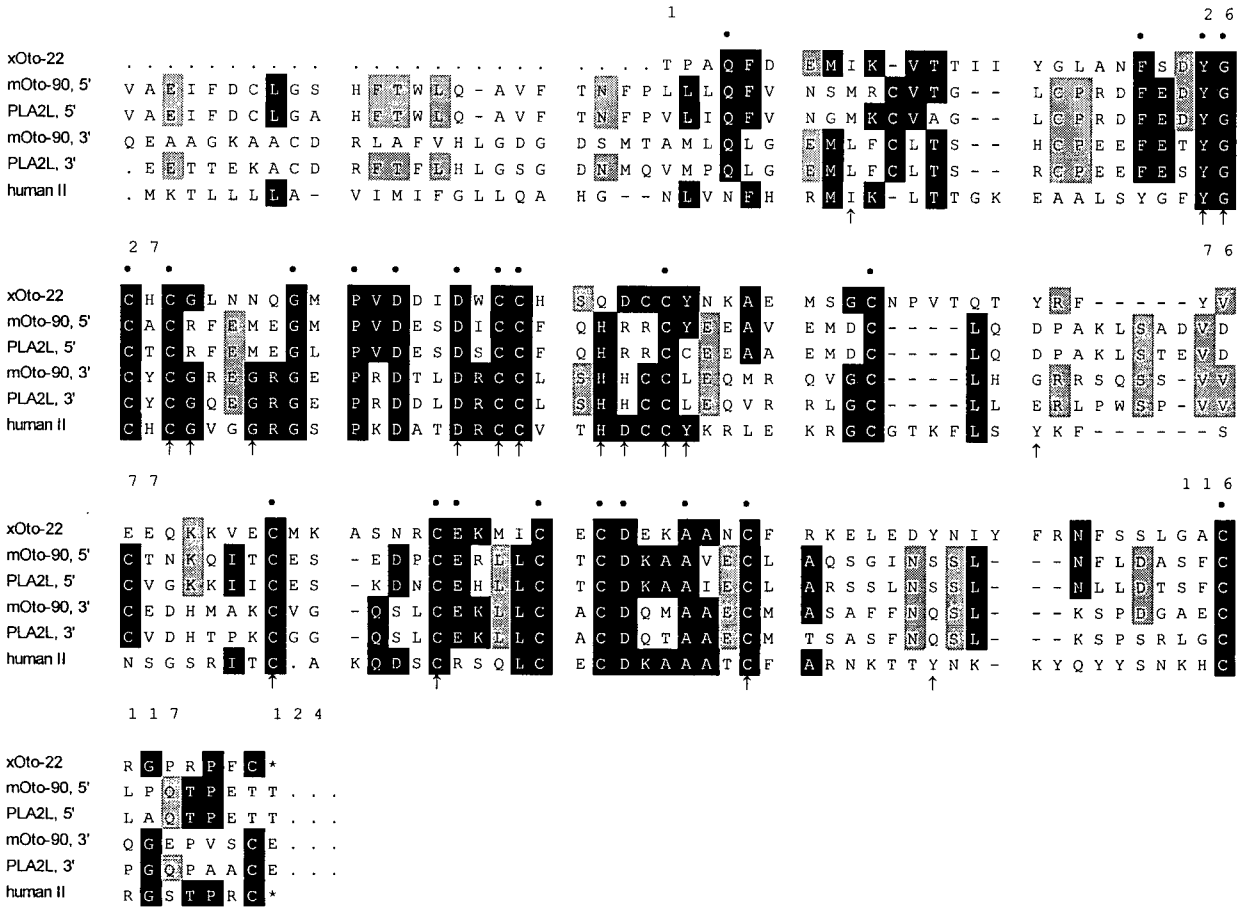


Figure 4.6 Alignment of otoconin PLA-domains

Alignment of amino acid sequences of the 5' and 3' PLA-domains of human (PLA2L) and murine otoconin, with the sequence of *Xenopus* otoconin-22, and human Group I $sPLA_2$. Residues that are conserved in at least 3 different sequences are shown in reverse print. Residues conserved in 3 of 5 otoconin PLA-domains are boxed in grey. Residues conserved among all otoconins are indicated by black circles above the column. Stop codons are shown as asterisks. Arrows indicate 18 conserved positions in all $sPLA_2$s. Numbering is according to the system of (Renetseder et al. 1985), position 1 is the start of both PLA2L-domains and mature, active $sPLA_2$. Dashes indicate gaps introduced for optimal alignment of all six sequences. The start of the PLA2L-domains of PLA2L is indicated with a 1. PLA2L5', 5' PLA-domain; PLA2L3', 3' PLA-domain; Human II, human Group II $sPLA_2$ (Seilhamer et al. 1989a); mOto-90, murine otoconin-90 (Chapter 4); xOto-22, *Xenopus* otoconin-22 (Pote et al. 1993). The amino acid lineup was generated using the GCG program Pileup followed by manual shading in Microsoft Excel.

mammalian otoconins occurred subsequent to *Xenopus* divergence from the mammalian lineage. However, the phospholipase A_2 paralog which makes up the *Xenopus* PLA-domain is, like those of its derivative mammalian PLA-domains, not closely related to any contemporary $sPLA_2$ gene, either mammalian or reptile Groups I or II. That the human PLA-domains were derived from a particularly ancient form of $sPLA_2$ has been shown by phylogenetic analysis (Feuchter-Murthy et al. 1993) and the *Xenopus* otoconin further emphasizes the age of the original gene duplication. Phylogenetic tree construction predicts the $sPLA_2$ paralog which composes the PLA-domain(s) branched from functional progenitor $sPLA_2$s prior to reptile radiation, i.e. very early in vertebrate evolution. The localization of both PLA2L/otoconin-90 to different chromosomes than functional human and murine $sPLA_2$ is further proof of this ancient origin.

Blocks of conserved sequence can be seen in Figures 4.5 and 4.6 with the most striking being the absolute conservation of cysteine residues between all otoconins, from human to frog, and $sPLA_2$. As $sPLA_2$ is known to be a heavily disulfide bonded, small, rigid protein, it appears that this characteristic is important for otoconin function as it is so strictly conserved, over the estimated 360 million years of evolutionary divergence between frog and human (Kumar and Hedges 1998). It is unknown how the formation of this small, rigid tertiary structure will be influenced by residing within the context of the non-PLA-domain sequences. Residues which are conserved in all otoconins, and therefore likely functionally important, are shown by black circles in Figure 4.6 and can be compared and contrasted to the residues conserved among all active $sPLA_2$ enzymes, which are marked with arrows.

Only 27% identity is found at the protein level between *Xenopus* otoconin-22 and murine otoconin-90. Some overlap between the two exists, primarily at the Cys residues, but also encompassing some others, such as Tyr^{25} and Asp^{39}, which are not thought to be important for secondary or tertiary structure. Perhaps the tyrosine residues are conserved because of essential phosphorylation; this conservation of residues required for post-translational modification is evident in the case of the N-linked glycosylation site, Asn-X-Ser/Thr (N-X-S/T). As the molecular weight of otoconin-90 predicted from the amino acid sequence is only 49 kDa, it is presumed to be quite heavily glycosylated (Wang et al. submitted),as is the truncated human PLA2L cDNA (Chapter 5). Four N-linked glycosylation consensus sites are found in the otoconin-90 cDNA (bold italics, Fig. 4.3) and of these, only Asn^{272} is not conserved between human and murine sequences, increasing the likelihood that the three sites represent authentic glycosylation sites.

4.2.4 Murine otoconin-90 has an independent and divergent 5' end

As otoconin-90 is known to be a secreted protein, and the N-terminal peptide sequence lacked a initiating methionine residue and signal sequence consistent of a protein cleaved by signal peptidase, it was realized that the cDNA sequence lacked a 5' UTR (untranslated region) and secretion signal sequence. The 5'UTR clone (Fig. 4.2) was generated using a PCR method to amplify the 5' end of transcripts present in a cDNA library. Briefly, iterative PCR using a vector primer and nested gene-specific primers was performed on a E14.5 murine cDNA library. The hybridizing products were cloned and 6 were sequenced; the consensus sequence consisted of 133 bp of upstream sequence, in frame to the rest of the transcript, and possessing a potential

initiating methionine preceding a 16 amino acid motif which was predicted to form a consensus secretion signal sequence. No larger hybridizing fragments were seen. As primer extension with MEST5 upon murine E14 RNA to map the start site of transcription was unsuccessful, it is impossible to be certain whether the actual 5' terminus has been cloned. The presence, however, of a predicted signal sequence at the N-terminus of a protein known to be secreted is reassuring. Interestingly, homology to human PLA2L ends immediately 5' to the murine signal sequence and 5' UTR (Fig. 4.4). This is predicted by my hypothesis, i.e. if otoconin-90 is the murine homolog of the downstream human otoconin gene in PLA2L, but is expressed from its native promoter and not a HERV-H LTR, the sequences directly downstream of this native promoter will diverge from that of PLA2L, which is expressed from a LTR promoter and *splices* into (at minimum) the 2nd exon of otoconin. This can be seen in Figure 4.10. This is due to the fact that the first exon of a gene, which contains the 5' UTR, lacks a splice acceptor sequence; therefore fusion transcripts generated by intergenic splicing cannot contain the first exon/5'UTR of a given gene.

4.2.5 Prediction of secretion signals in analogous regions of murine otoconin-90 and human PLA2L

Murine otoconin-90 was found to be the primary protein component of otoconia by direct amino-terminal (N-terminal) microsequencing. As otoconia are extracellular bioorganic crystals within the inner ear, otoconin-90 protein appears to be secreted from inner ear non-sensory epithelial cells (Wang et al. submitted). The sequenced N-terminal peptide lacked an initiating methionine, which is indicative of post-translational cleavage of the nascent protein. Very likely the initiating Met and

subsequent signal sequence was removed by signal peptidase cleavage when the pre-otoconin-90 was translocated into the lumen of the endoplasmic reticulum (ER).

Signal peptides target proteins to the secretory machinery, and while there is little sequence similarity between various signal sequences, common features exist. This has been most strikingly shown by the ability of many different signal sequences to enable secretion of their associated proteins, and by the ability of signal sequences from other organisms to direct secretion of a heterologous gene in divergent organisms (Zheng and Gierasch 1996). The deduced common structure consists of a N-region containing a positive charge, a hydrophobic H-region followed by a polar and neutral C-region (Nielsen et al. 1997). Additionally, it has been found that the residues at positions -3 and -1 relative to the cleavage site must be neutral and small for correct signal peptidase cleavage to occur; this is termed the -3, -1 rule (von Heijne 1985).

Characterization of the murine 5' UTR clone (Fig. 4.2) revealed a possible initiating methionine codon followed by 16 amino acids, in frame to the rest of otoconin-90 ORF. Notably, there were no additional Met codons in the same frame upstream, but there were also no in-frame stop codons. This amino acid sequence was analyzed by the SignalP program (*http://www.cbs.dtu.dk/services/*) which uses combined artificial neural networks, with one trained to recognize the cleavage site and a separate one to differentiate between signal peptides and non-signal peptides. The widely used weight matrix method (von Heijne 1986) is used to determine the cleavage site, and the dataset used to train these networks consisted of 416 secreted human proteins, and 1011 eukaryotic proteins. The SignalP program's performance is as follows: it can accurately distinguish between signal and non-signal sequences (R = 0.97) while the cleavage site prediction is not as accurate (68% correct) based on the

human dataset (Nielsen et al. 1997). As shown in Figure 4.7A, the SignalP program predicted the presence of a signal peptide at the N-terminus of the otoconin-90 protein, and notably predicted the cleavage site to occur between the A and H residues (...VGA*HALD....) which corresponds with the observed N-terminus of the cleaved, mature otoconin-90 protein (HALD....). This sequence also adhered to -3, -1 rule, by possessing small neutral valine and alanine residues at the requisite sites relative to the cleavage site. It should be noted that the program did not locate any further signal peptides when given other length-matched regions of otoconin-90, including sequences encoding part of the predicted signal sequence. The graphical output of the SignalP consists of a three-variable graph (Fig. 4.7), with relative score on the Y-axis and the protein sequence and position on the X. The C-score, or raw cleavage site score is expressed as a bar graph in green, and will be high immediately after the cleavage site (+1). The signal peptide score, or S-score, shown as a blue dashed line, will be high at all positions before the cleavage site and drops sharply *at* the cleavage site. It is low at non-signal peptide N-termini. The Y-score, expressed as a red dotted line, is the combined cleavage site score; the optimized prediction of cleave site location. This is achieved by combining the height of the C-score with the slope of the S-score, to determine where the C-score is high and the S-score drops radically from high to low values. An example of this can be seen in Figure 4.7A, where two C-score and Y-score peaks of equal heights can be seen, but the correct cleavage site is predicted due to the higher slope of the S-score as it intersects the first (and correct) C-peak. This result was confirmed by using a similar program using a different search algorithm (*http://bimas.dcrt.nih.gov:80/molbio/signal/*).

When the region homologous to the murine signal peptide was examined in PLA2L, only 41% identity (and 76% similarity) was seen at the amino acid level, over the 17-mer peptide. Interestingly, this sequence contained a possible initiating methionine residue at the identical position relative to the murine signal sequence (see Fig. 4.8), and small, neutral amino acids alanine and glycine at the -3,-1 positions. When this sequence was analyzed by SignalP, as shown in Figure 4.7B, it was predicted to form a signal sequence, and to be cleaved prior the H residue (...AGG*HPLD....). The finding that the PLA2L transcript contains a predicted signal peptide and putative initiating methionine at the 5' end of otoconin-90 homology lends further support to the hypothesis that PLA2L represents a HERV-H-induced fusion of two independent but adjacent transcriptional units, where the latter unit encodes the human ortholog of otoconin-90. As no teratocarcinoma-specific chromosomal rearrangement in the PLA2L region exists ((Feuchter-Murthy et al. 1993), and D. Mager, unpublished results), this fusion must have necessarily been caused by intergenic splicing, with the site of fusion occurring 46 bp upstream of the signal sequence, in the original PLA2L cDNA. Additional support for this hypothesis resulted from the genomic cloning of the interval encompassed between the two exons on either side of the fusion (exons 2 and A2 in Fig. 4.10, or 4 and 5 in Fig. 4.9A).

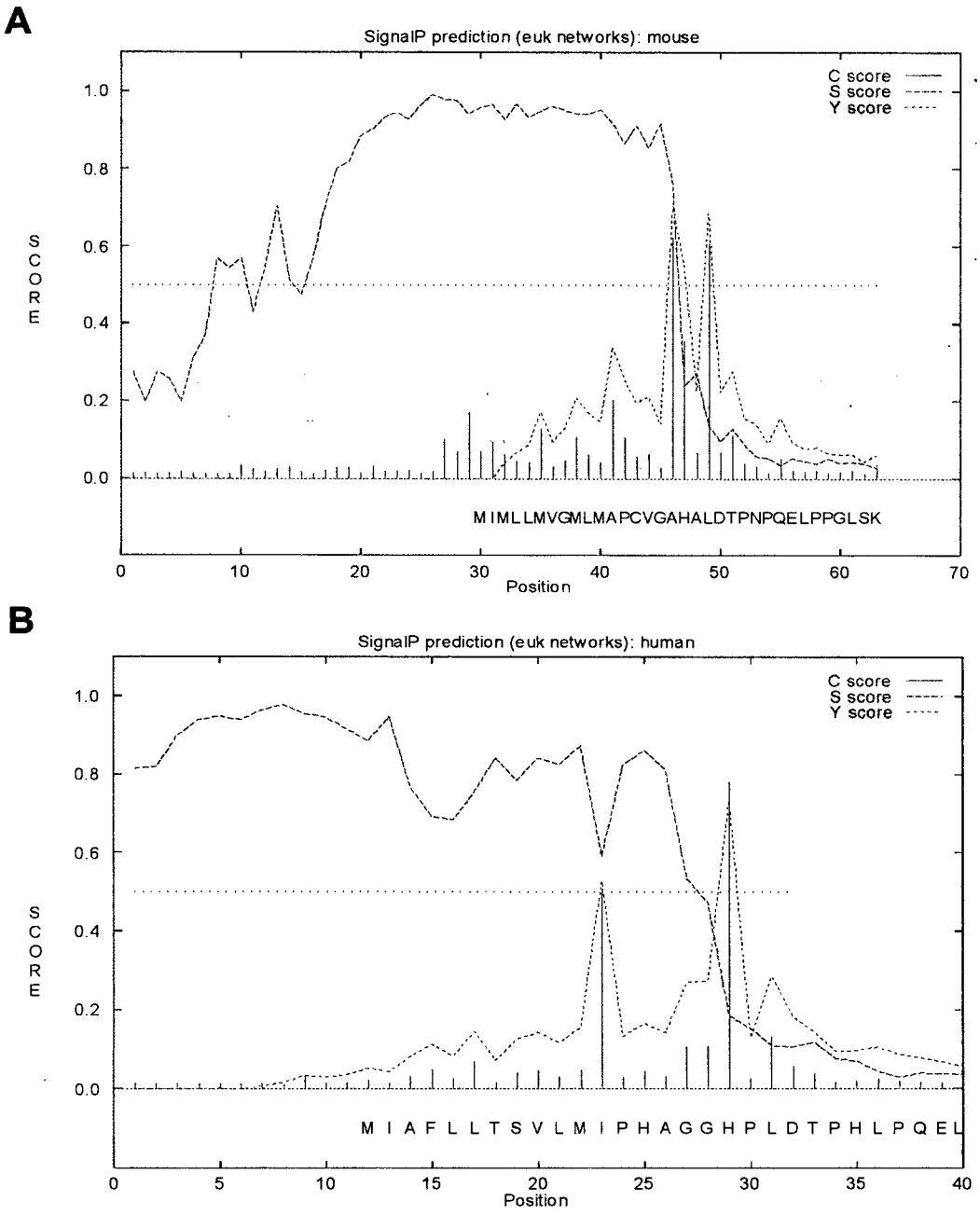


Figure 4.7 Predicted secretion signals in otoconin-90 and PLA2L
The 17-mer in-frame amino acid sequence found in the murine 5' UTR clone was predicted to form a signal sequence **(A)** by the SignalP program, using the weight matrix method of von Heijne. The cleavage site is predicted to occur between the A and H residues (VGA*HALD) which is accurate, as shown by peptide sequencing. When the analogous sequence from PLA2L was used, which was matched on the position, the length and the presence of a initiating Met residue, the SignalP program predicted a signal sequence, with a cleavage site in the same position **(B)**. The human and murine signal sequences used were identical at only 7/17 residues, and similar at 5. The method used to predict signal sequences is discussed in the text.

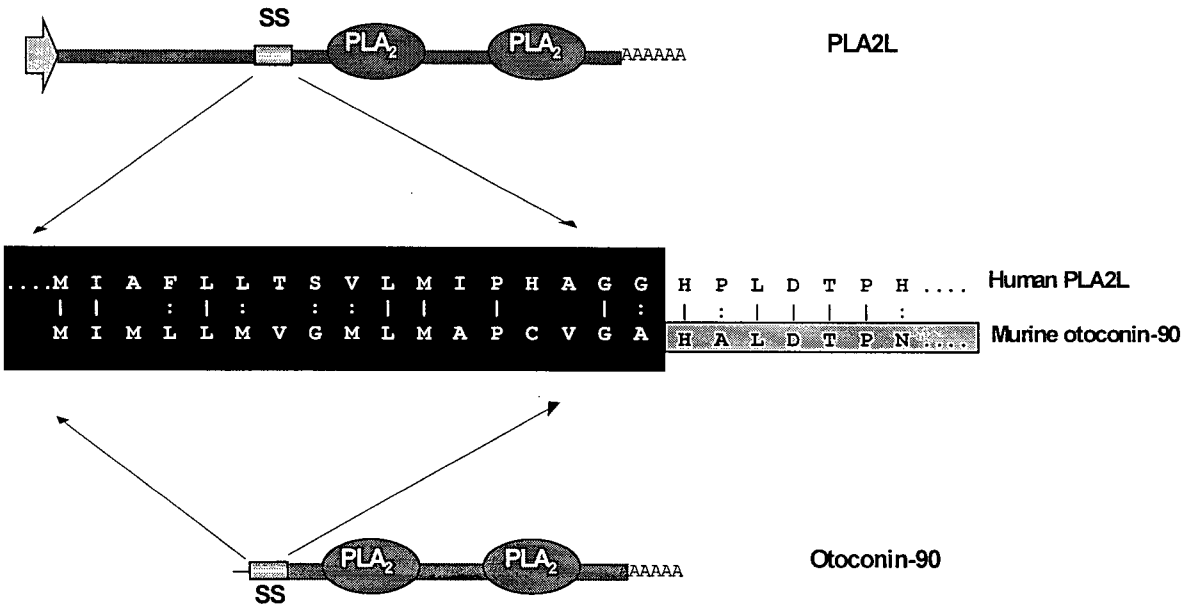


Figure 4.8 Localization of signal peptides in PLA2L and otoconin-90 proteins
Predicted secretion signal peptide locations within each protein shown diagrammatically, human PLA2L at top and murine otoconin-90 at bottom. Peptide sequence and alignment in reverse print indicates complete signal peptides. Peptide sequence boxed in grey shows the N-terminus of the original murine protein. SS: signal sequence.

4.2.6 Cloning of human intergenic genomic region

When murine otoconin-90 was determined to be a homologous independent gene without the upstream exons seen in PLA2L, examination of the human genomic region directly upstream of the start of otoconin-90 homology was undertaken. It was initially assumed that an intron, relatively small like most PLA2L introns (but containing 2 small exons, termed 3 and 4), was present between the exon 2 and exon 5, as shown in the AF-5 cDNA in Figure 4.9A, and in Figure 3.2 (shown as exon 2 and exon A2). Surprisingly, genomic cloning of this region showed it was much larger than expected, with a final size of approximately 30 kb. This region was cloned in a contig using long-range genomic PCR, with primers derived either from known exonic sequence or newly-sequenced intron sequence, as detailed in Materials and Methods and Chapter 3. As discussed in Chapter 3, elucidation of the genomic organization of this region revealed that the transcribed sequence (5R2.1), found at the 3' end of the AF-6 and 7 PLA2L cDNA clones and previously thought to be located *downstream* of the PLA-domain exons, was actually located *upstream* of the PLA-domains. This finding suggested that the structure of the PLA2L transcript was more complex than originally believed. The original model of the structure of the PLA2L transcript, including the AF6/AF7-derived 5R2.1 sequences not found in the prototypical AF-5 cDNA clone, is shown in Figure 4.9a. The cloning and characterization of what was initially seen as a small intron between exons "2" and 5 (shown as a star in Fig. 4.9a) revealed an approximately 30 kb genomic gap, which contained the 5R2.1 sequences. This finding led to a revision of the order of exons and cDNA clones relative to the genomic map, and, combined with studies of the murine homolog, to the conclusion

that the genomic gap is an intergenic space between two adjacent unrelated genes, which compose the fusion transcript PLA2L.

These important shifts in the knowledge of PLA2L locus structure are shown in Figure 4.9B, with the primary finding being that the upstream exons are separated from the PLA-domain exons by a larger genomic distance than originally thought. Additionally, the positioning of the 5R2.1 region at the 3' terminus of the upstream exons was both surprising and interesting, as the 5R2.1 region has all the characteristics of a 3' UTR. The 5R2.1 region is a 2.1 kb transcribed sequence (isolated as a cDNA clone) which lacks an open reading frame and is likely coded by a single exon, all of which are known features of 3' UTRs: length, noncoding and lack of introns (Berget 1995). These results, in concert with the previous findings regarding the mouse homologue, its divergent 5' end, and predicted signal peptides have led to a revision of PLA2L structure. It is now hypothesized that PLA2L is a HERV-H-associated tripartite fusion transcript composed of HERV-H sequences, a number of exons of a anonymous but conserved gene termed HHAG1 (HERV-H Associated Gene 1) which is then intergenically spliced to the 15 exons of the human otoconin-90 gene. These two genes lie in the same orientation in human genomic DNA, and very close together; only ~10 kb separates the end of the HHAG1 3' UTR from what is likely the *second* exon of human otoconin-90 (Fig. 4.10). Although variable, this gene density is very different from the predicted average human intergenic distance of ~50 kb (B. Roe, personal communication), and this proximity may in fact be a possible explanation for the necessary transcriptional antitermination event prior to intergenic splicing.

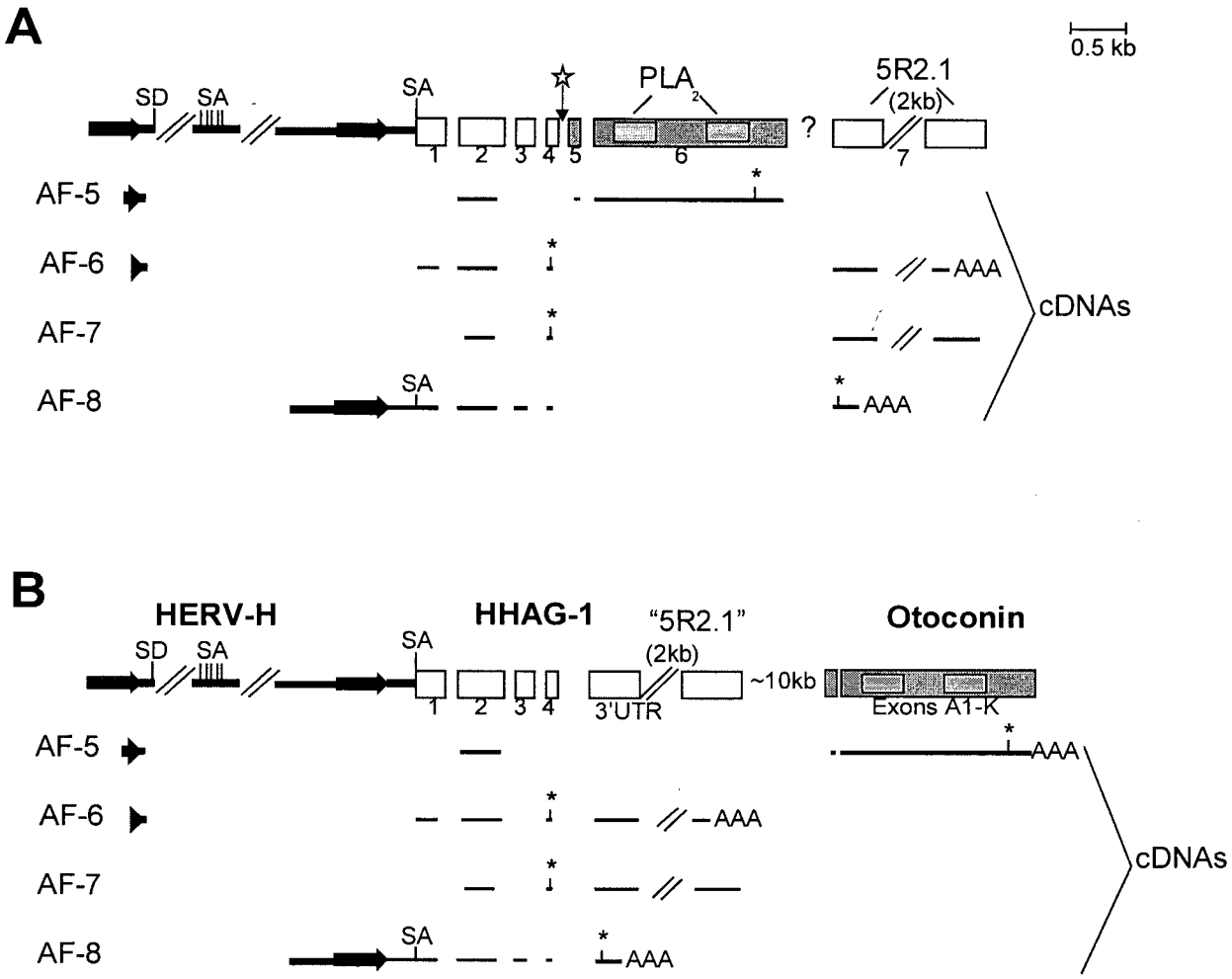


Figure 4.9 Original and revised composite PLA2L structure
The composite structure of the PLA2L transcript and locus was originally deduced by comparing and partially sequencing the 4 originally isolated cDNA clones (AF clones) and a number of Tera1-derived RT-PCR clones (not shown). Only one clone, AF-5, contained the PLA-domain exons, while 2 of the other 3 contained a large 2 kb sequence at the 3' terminus (5R2.1 or 7), which was not found in AF-5. The original view of the structure with its aligned cDNAs below is shown in **(A)**, modified from (Feuchter-Murthy et al. 1993). Subsequent human genomic and murine homolog cloning have elucidated the genomic position of the "exon" 7 in panel A, or 5R2.1 in panel B. Based upon current knowledge, **(B)** shows a schematic of the PLA2L region, which is composed of the HERV-H element (shown in thick line), 8-10 exons of an unknown gene termed HHAG1 (HERV-H Associated Gene), and the 14 exons of the human otoconin gene. SD; splice donor, SA; splice acceptor, *; stop codon. The star symbol in panel A represents the site of fusion by intergenic splicing, and the intergenic genomic region. HHAG1 "exons" 1-4 do not represent actual exons but groups of exons in some cases. Exon A2 corresponds to exon 5 (grey) in panel A.

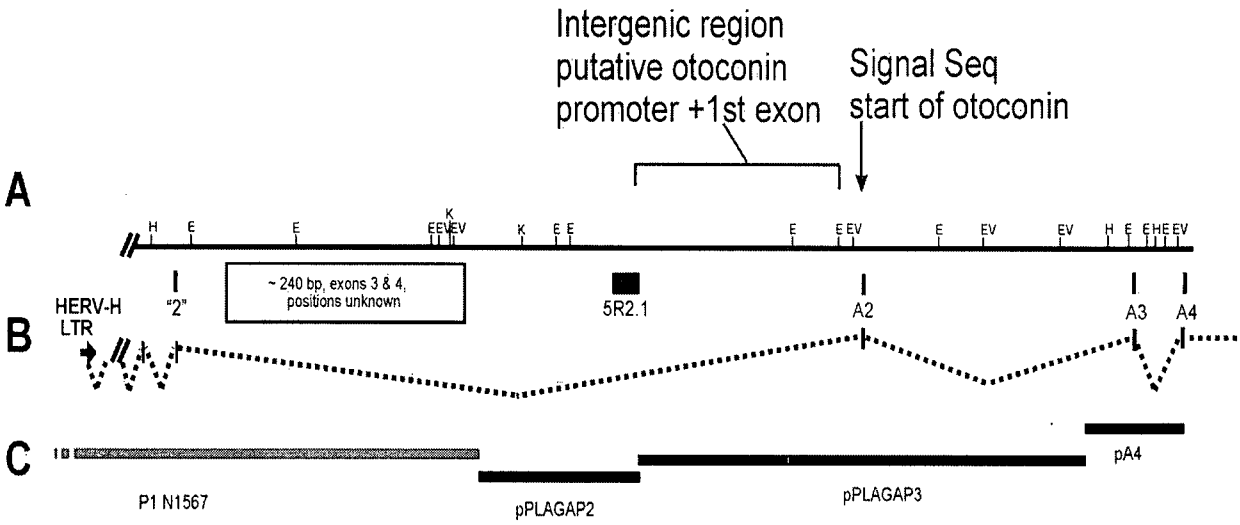


Figure 4.10 Detailed schematic map of human intergenic region

This figure is an expansion of the intergenic region in Fig. 3.2. A) HHAG-1 / PLA2L intergenic genomic DNA is shown as a black line. HHAG-1 exon "2" is shown below the line, as is the exon it is fused to in the AF-5 cDNA, A2 of otoconin-90. Large exon which likely encodes the HHAG-1 3' UTR is also shown below the line, as 5R2.1. Intergenic space (which contains the first exon of otoconin-90) shown bracketed above the line, as is the signal sequence-containing exon A2. Parallel diagonal lines at left indicate location of HERV-H not to scale. B) Structure of original AF-5 PLA2L cDNA, showing intergenic splicing between HHAG-1 exon "2" and otoconin-90 exon A2. Exon A2 corresponds to "exon 5" in Fig.4.9A. Terminal exons "5R2.1" and A1 were not included as they lack the requisite flanking splice sites. C) Genomic clones used to determine intergenic genomic region.

4.3 Discussion

4.3.1 Otoconin-90 expression

The human PLA2L transcript was initially thought to be the product of a single gene, which was being expressed from a heterologous HERV-H LTR promoter. This model was thought to be accurate due to PLA2L possessing a number of characteristics of a solitary gene, the primary one being the mapping of the cDNA to a single genomic locus. In addition, the presence of a long ORF in the original cDNAs and the lack of any detectable LTR-independent expression further reinforced this belief. As shown in Figure 4.9A, cDNAs containing only the anonymous upstream domains were also isolated, and the lack of PLA-domains was explained by alternative splicing or polyadenylation. This hypothesis was called into question with the realization that the homologous murine gene, otoconin-90, contained only the PLA-domains and associated proximal exons, and lacked the upstream exons found in PLA2L. The genomic cloning of the interval between the two genes and the positioning of what is very likely the 3'UTR of HHAG1 (5R2.1 in Fig. 4.9) at the 3' terminus of that gene strongly suggests that the PLA2L cDNA was not derived from a single transcriptional unit. The eventual cloning of the otoconin-90 5' UTR and finding it divergent from anything present in the human upstream exons was the convincing evidence that the PLA-domain containing otoconin-90 was being expressed as an independent gene, without the anonymous upstream exons inPLA2L.

Over 50 human tissue and cell line RNAs were screened by dot-blot hybridization, northern blotting or RT-PCR to detect evidence of PLA2L expression. With the exception of the two independent teratocarcinoma lines Tera1 and

NTera2D1, no expression was seen, and no PLA2L human ESTs have ever been sequenced, despite over one million human ESTs from 650 cDNA libraries being present (as of May, 1998). This paucity of expression can be explained by the extremely small window of temporal and spatial expression exhibited by the murine otoconin-90 gene, and, by extrapolation, human otoconin-90. Interestingly, no evidence of HHAG1 expression has been seen in human, outside of teratocarcinoma RNAs. However, the presence of multiple alternatively spliced exons with an ORF, and evidence of conservation in mouse (D. Mager, unpublished observations) imply that it is a *bona fide* gene. The normal expression of HHAG1 is of interest, as the HERV-H insertion occurred within this gene, and it is far likelier to be affected by the HERV-H element than is human otoconin-90. It is provocative to speculate that HHAG1 and otoconin-90 may have similar inner-ear-specific expression patterns, due to their close genomic proximity. Genomic clustering of genes of similar function or expression is well known, some examples include the complement gene cluster and the MHC gene cluster (Ashfield et al. 1994).

Murine otoconin-90 ESTs were isolated from E14 mouse embryo libraries, and in situ hybridization subsequently showed expression in the embryonic otocyst, localized to the non-sensory epithelia, from embryonic days E10 to E17 (Wang et al. submitted). Notably, expression was not detected in the adult inner ear, and it was seen only in a subset of epithelial cells in the embryonic inner ear. This restricted expression pattern provides an explanation for the lack of human otoconin-90 expression seen; no human whole-embryo cDNA libraries exist, and although 2 human ear-derived libraries were used for EST generation, neither were derived from the fetal inner ear. In fact, no human fetal inner ear cDNA library has ever been reported in the

literature, as access to fresh fetal tissue from the appropriate stage would be extremely difficult to obtain. This is the reason that human otoconin-90 expression, independent of HERV-H LTR expression, has not been examined.

This extremely restricted expression pattern is not unique, as many important developmentally regulated genes are only transcribed at certain periods in certain small populations of cells. This is thought to be due to these genes possessing complex promoters which are only responsive to specific transcription factors, or to the complex interaction of multiple factors, which are themselves expressed in a very restricted manner. Examples of this phenomena in the inner ear include the inner ear-specific saccular collagen gene (Davis et al. 1995), the Ocp2 transcription factor expressed only in the cochlea (Chen et al. 1995), and the Brn-3.1 transcription factor which is exclusively expressed in embryonic cochlear hair cells (Erkman et al. 1996). In addition, there are examples of genes with somewhat wider expression patterns, whose homologous knockout appears to affect only the inner ear, such as the IsK potassium channel gene (Vetter et al. 1996).

4.3.2 Structural implications of conserved PLA-domains

The primary structural module present in the three known otoconin genes is the PLA-domain, and is very interesting from a functional viewpoint. The absolute conservation, between human and *Xenopus*, of the cysteines making up the disulfide bridges implies that the rigid $sPLA_2$ conformation is functionally relevant. Otoconin function is unclear, but as it is the primary protein component of otoconia, it suggests that it is a structural protein involved in otoconia biosynthesis. Otoconia crystallize from calcium carbonate, so it is likely that otoconin-90 plays a important role in calcium binding and crystallization. On the surface, a $sPLA_2$ structure seems ideal for a protein

involved in calcium binding, as phospholipase A_2 enzymes require and bind Ca^{2+}. When examined carefully, however, it is seen that the substitutions in what would be the enzyme's active site (Asp^{49}) will also result in abrogation of Ca^{2+} binding. As $sPLA_2$ is one of the most structurally characterized proteins, it is known that the calcium ion is coordinated and bound by the Asp^{49} and the His/Tyr^{28}, Gly^{30} and Gly^{32} triad (Scott et al. 1991). Both mammalian PLA-domains are substituted for the Asp^{49}, but interestingly the 2nd PLA-domain in human and mouse otoconin retain the rest of the calcium binding residues. Crystallography of mutated $sPLA_2$ enzymes has shown that any change in the 4-residue calcium binding cage results in the abrogation of calcium binding (Li et al. 1994), with the exception of a recently described catalytically active, calcium-binding viper $sPLA_2$ which contains a Ser^{49} in the place of Asp^{49} (Polgar et al. 1996).

Snake venom has long been known as a rich and complex source of phospholipase A_2 enzymes, and in fact they make up the primary toxic component. Interestingly, various *Viperidae* snake venoms contain well-characterized inactive $sPLA_2$s whose lack of catalytic activity is due to a Asp^{49} substitution very similar to that of the otoconin PLA-domains. These inactive homologs are termed K^{49} or Lys^{49} PLA_2s, due to their substitution of the essential acidic Asp^{49} with the basic Lys^{49}, and function solely as rapidly acting myotoxins (Selistre de Araujo et al. 1996). These myotoxins do not possess phospholipolytic activity due to the abrogation of Ca^{2+} binding, which is caused by the combined steric hindrance and improper charge of the long and positive lysine side-chain, which protrudes into the space usually occupied by the calcium ion (Scott et al. 1992). This is analogous to the substitutions in the mammalian PLA-domains, which are both positively charged and sterically bulky, His^{49} or Arg^{49}, and

guarantee the lack of both PLA_2 activity and direct Ca^{2+} binding in the otoconin PLA-domains. Due to the conservation of cysteine disulfide bond pairs and the fact that the crystal structures are known for many $sPLA_2$ enzymes, the tertiary structure of the otoconin PLA-domains can be readily predicted by computer superimposition of the domains onto the backbone of a known $sPLA_2$, as was done for *Xenopus* otoconin-22 (Pote et al. 1993). This would be of considerably less value for the mammalian otoconins, and it was not attempted for that reason, as the relevant question cannot be answered using this procedure; i.e. how do the PLA-domains fold within the *context* of the rest of the otoconin protein? Very likely the PLA-domains will adopt similar structures to $sPLA_2$, but how the remainder of the protein would affect the structure of these domains cannot be modeled. This is due to the non-PLA-domain regions of otoconin lacking similarity to any known protein, structurally characterized or not, resulting in no "backbone" to use as a predictor. *Xenopus* otoconin-22, which structurally consists of only a single PLA-domain and lacks the extensive leader, intradomain and tail regions seen in mammalian otoconin-90 proteins, has been shown to lack PLA_2 activity and does not bind radioactive ^{45}Ca (Pote et al. 1993). The domains may not bind calcium via the PLA-domains, but may provide a template for crystal growth via complexing calcium carbonate with aspartate residues, as shown in molluscan shell biogenesis systems (Wheeler 1992).

The specific conservation, over hundreds of millions of years, of the sPLA2 Cys:Cys disulfide bond positions in otoconin PLA-domains coupled with the specific lack of conservation of the phospholipase catalytic site residues is important to note (Fig. 4.6), as this implies that the PLA-domain modules were duplicated and retained within the larger otoconin protein by virtue of their *tertiary* structure. The bulk of

conserved protein domains and motifs are conserved on the basis of function, or amino acid (primary) structure, such as the DNA-binding C_2H_2 zinc finger motif or the phosphotyrosine-binding SH2 domain. Often this primary sequence forms a desired secondary, or occasionally tertiary, structure (Henikoff et al. 1997). It is much more uncommon to find a sequence conserved, not by virtue of its functional or catalytic role, but for the manner by which it folds, its tertiary structure. The primary example of this conservation-for-tertiary-structure event is the α-crystallin domain found in lens crystallins, small heat shock proteins, and other diverse proteins, including mycobacterial surface antigens (de Jong et al. 1993). It is thought that the tertiary structures of these domains are very stable and long-lived, which is likely the basis of their conservation and diverse usage. The origin of these domains is controversial, but it appears they were recruited from small heat shock proteins (Piatigorsky 1990). Interestingly, and reminiscent of the otoconins, another possible reason for the long conservation of α-crystallin domains is their ability to accumulate to very high concentrations without precipitating, with greater than 50% of eye lens protein of various vertebrates composed of α-crystallin (de Jong et al. 1993).

4.3.3 Antitermination and intergenic splicing

The mechanism by which two genes are fused by splicing, in the absence of gross chromosomal rearrangement, is unclear. With the possible exception of two examples, most of reported cases of intergenic splicing and subsequent gene fusion occur between oncogenes in leukemias, and are the result of chromosomal rearrangements juxtaposing exons from one gene adjacent to a breakpoint, next to breakpoint-proximal exons of a distant gene (Nucifora et al. 1994). The first possible example of intergenic splicing causing fusion of two normally adjacent genes is that of

the human POM-ZP3 bipartite transcript (Kipersztok et al. 1995). This novel mRNA is found in a number of human tissues, and appears to be a fusion of an upstream human homologue of the rat nuclear pore protein POM121 to the downstream exons of a ovum zona pellucida protein, ZP3. The authors show both regions of this bipartite transcript to be localized to the same region of human chromosome 7q11, but do not present any genomic cloning data supporting intergenic splicing. Additionally, the authors of this study speculate that the simplest explanation for formation of the bipartite transcript is duplication and juxtaposition of the relevant genomic fragments. No data suggesting this event is shown or proposed, and the lack of knowledge of this locus, known only from a single publication, weakens it as an example of intergenic splicing. In contrast, the second example is well characterized and represents an authentic intergenic fusion by splicing; between a gene of ubiquitous expression but unknown function, MDS1 and an adjacent transcriptional repressor, Evi1 (Nucifora et al. 1994). This fusion transcript is known from myeloid leukemias, where breakpoint-fusions to the Evi1 oncogene are common. Uniquely, the MDS1/Evi1 fusions are seen in the *absence* of chromosomal rearrangements, and in normal tissues, even when examined at the molecular level (Fears et al. 1996). Independent expression of both the MDS1 gene and the Evi1 gene has been shown using a number of methods, indicative of their status as independent transcriptional units. Similar to experiments performed to prove HHAG1 is independent of otoconin-90 in human, cDNAs were isolated showing MDS1 expression without Evi1 sequence, and poly(A)signals in the 3' UTR of MDS1 were identified and shown to be functional. Additionally, as both genes are widely expressed (contrasting HHAG1/otoconin-90 expression), RNase protection and primer extension were performed to further delineate the independent

expression of MDS1 and Evi1 (Fears et al. 1996). The genes are located adjacently on human chromosome 3q26, and are surprisingly ~170 kb distant from each other, as shown by pulsed-field mapping, and the region is not rearranged in cells where MDS1/Evi1 fusion expression is seen (Nucifora et al. 1994; Nucifora 1997). The MDS1/Evi1 fusion creates a PR-domain-containing strong transcriptional activator, which is oncogenic (Soderholm et al. 1997). This example seems to represent an authentic precedent for intergenic splicing resulting in gene fusion, with the splicing occurring over a far greater intergenic region than that of HHAG1/otoconin-90 (10 kb).

Eukaryotic intergenic splicing appears to be a very rare event, in the absence of chromosomal rearrangements, and its mechanism is essentially unstudied. This is in contrast to the well-known examples of bacterial catabolic enzymes. However, eukaryotic splicing and the linked event, transcriptional termination, is extensively studied, and implies a mechanism for gene fusion via intergenic splicing. As *trans*-splicing *in vivo* is apparently restricted to lower eukaryotes such as *C. elegans*, the manner by which two adjacent transcriptional units are fused is by loss of transcriptional termination/polyadenylation, or antitermination, followed by transcription of the entire intergenic region into pre-mRNA, and subsequent removal of the intergenic region, and joining of exons derived from separate genes, by splicing. It is a testament to the processivity of mammalian RNA polymerase II (RNAP II), that it commonly transcribes hundreds of kilobases of intronic DNA into pre-mRNA without apparent detriment – energetic or otherwise- when genes containing many large introns are transcribed. Although others exist, the primary model for eukaryotic termination predicts that the poly(A)site cleavage reaction and subsequent 5' → 3' exonucleolytic degradation of the nascent end of the pre-mRNA is the initiating step for

the termination of transcription; this model has been recently proven in yeast (Birse et al. 1998). This model therefore predicts that defective termination and RNA 3' end-processing could result in intergenic fusions. If the poly(A)signal in the 3' UTR of HHAG1 is weak by sequence-specific reasons or if polyadenylation is suppressed by the U1 SnRNP proteins A1 or 70 kDa (which are bound to the final splice acceptor site) inhibiting poly(A) polymerase (Gunderson et al. 1998), cleavage of the HHAG1 pre-mRNA could be suppressed, leading to loss of transcriptional termination and resulting in transcriptional read-though of the HHAG1/otoconin-90 intergenic region (see Fig. 4.10 for the intergenic region and putative site for the otoconin promoter) and incorporation of otoconin-90 exons into the tripartite PLA2L transcript seen. Notably, the original fusion cDNAs AF-6, AF-7 and AF-8 are all polyadenylated at different locations within the 3' UTR (5R.2.1, Fig. 4.9B) providing circumstantial evidence for a weak HHAG1 poly(A)site. Alternatively, when a poly(A)signal is weak or genes are closely spaced a secondary DNA element, downstream of the poly(A)site, is needed to ensure efficient transcriptional termination and to prevent promoter occlusion of the downstream gene. The 10 kb between HHAG1 and otoconin-90 is considered closely spaced. This DNA element, called a terminator sequence, is usually a motif recognized by a DNA-binding protein which binds and then bends the DNA, resulting in a "roadblock" for a elongating RNA polymerase. The only known eukaryotic RNAP II terminator sequence is G_5AG_5, found between the closely spaced genes human complement C2 and Factor B genes, human MHC Class III genes g11 and C4, and murine IgM and D genes, which serves as a high-affinity binding site for the MAZ transcription factor (Ashfield et al. 1994; Moreira et al. 1995). This protein or a similar "terminator", may be lacking or limiting in the teratocarcinoma tumor cells which PLA2L

was derived from, or possibly a RNAP II elongation stimulator such as TFIIS may be overexpressed, resulting in a slightly favorable environment for intergenic fusion. It should be noted that the intergenic splicing which joins the HHAG1 exons to the otoconin-90 exons to form the PLA2L transcript in human teratocarcinoma cell lines may be teratocarcinoma tumor cell – specific event, but it is not a rare or random artifactual event in these cell lines. Multiple cDNA and RT-PCR clones from 2 unrelated teratocarcinoma lines (Tera1 and NTera2D1) have been isolated, all showing this fusion via splicing. This fusion transcript was detected by virtue of the uniquely high promoter activity of HERV-H LTRs in teratocarcinoma cell lines. It is tempting to ascribe some connection between the integration of an endogenous retrovirus with a transcriptionally active LTR, and gene fusions. However, there are no data to support this. The exact mechanism by which two adjacent genes are spliced together, and why this event appears to be favored in teratocarcinoma cells, awaits elucidation.

CHAPTER FIVE: HERV-H SUPPRESSES TRANSLATION OF AN ASSOCIATED FUSION TRANSCRIPT, PLA2L

A majority of the data presented in this chapter composed the following manuscript:

Kowalski, P.E. and Mager, D.L. A human endogenous retrovirus suppresses translation of an associated fusion transcript, PLA2L. J. Virol. 72: 6164-6168, 1998.

5.1 Introduction

Endogenous retroviruses can potentially affect cellular genes by promoting deletions or translocations due to inter-element recombination, or by retrotransposition. Depending upon the location, HERV insertions have the potential to cause such alterations in gene expression as alternate tissue specificity, inappropriate promoter activity, premature truncation of a reading frame via introduction of a frameshift or nonsense codon, or alternate polyadenylation (Liu and Abraham 1991; Goodchild et al. 1992; Ting et al. 1992; Amariglio and Rechavi 1993; Wilkinson et al. 1994; Di Cristofano et al. 1995; Schulte et al. 1996). By far the most common effects exerted by ERVs upon cellular genes are at the transcriptional level, not that of translation. Because the insertion of a ERV into a transcriptional unit is usually a very disruptive event, the damage is often realized at the immediate level of transcription, whereas to effect a translational alteration, the damage must be much more subtle while allowing transcription to occur.

As discussed in previous chapters, a HERV-H LTR promotes the transcription of a fusion transcript, termed PLA2L, which contains a short segment of LTR and leader region of HERV-H sequence spliced to downstream exons. As this transcript possesses HERV-derived sequences at its 5' terminus, and 5' untranslated regions (5' UTRs) of vertebrate genes are known to regulate the initiation of translation (Sachs and Buratowski 1997), the regulation of protein synthesis of PLA2L was studied. Here I report that HERV-H sequences, acting as a 5' UTR, serve to suppress translation of the PLA2L transcript in both the original teratocarcinoma cell line and in a heterologous expression system.

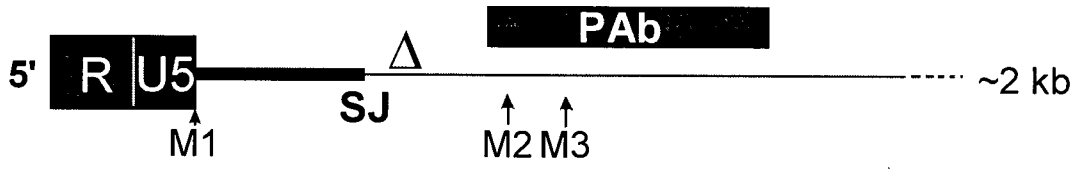
5.2 Results

Figure 5.1 shows the 5' region of the PLA2L transcript, where the HERV-H / PLA2L fusion occurs. In the 5' terminal 500 bp of the transcript, there are three potential initiating methionine residues within the same reading frame, which possess Kozak consensus sequences of varying quality. The context of the first two AUG codons (nt 101 and 416, respectively) is suboptimal, especially with regard to the crucial -3 position, which is a purine in almost all true initiating codons (Kozak 1996). The last AUG codon (nt 455) matches the Kozak consensus to a much greater extent, including both the A^{-3} position and the G^{+4}, and it is considered to be the most likely initiation codon for the PLA2L fusion transcript.

5.2.1 Expression and purification of PLA2L fusion proteins

To determine the regulatory role that the HERV-H element may play at the PLA2L locus, polyclonal antiserum against PLA2L was raised. Two PLA2L:glutathione-s-transferase (GST) fusion proteins, PLA2L-AF3 and pGEX-PLA, were generated by PCR amplifying regions of the original PLA2L AF-5 cDNA. Both fusion proteins were used to generate rabbit polyclonal antisera, however the first, PLA2L-AF3, was the antiserum used in all subsequent experiments. To create PLA2L-AF3, bases 391-584 of the PLA2L cDNA were amplified and cloned into the Sma I site of pGEX2T (Pharmacia), in-frame to GST. Induced bacteria containing this construct were lysed by sonication. Affinity chromatography with glutathione-agarose beads (Sigma) was performed upon the cleared lysate. A Coomassie Blue-stained SDS polyacrylamide gel of the PLA2L-AF3 protein can be seen in the left panel of Figure 5.2, with the an expected molecular weight of 37 kDa. Equal amounts of bead-bound and competed fusion proteins were loaded. As detailed in the Materials and Methods chapter, the

A



100 bp

	-3 +1 +4
Consensus	CRCCATGG
M1	gcgCATGa
M2	aGCCATGa
M3	gAtaATGG

B

```
  1  GAATTCCCTG ACTCTCTTTT CGGACTCAGC CCGCCTGCAC CCAGGTGAAA
 51  TAAACAGCTT TATTGCTCAC ACAAAGCCTG TTTGGTGGTC TCTTCACACG
101  GACGCGCATG AAATTTGGTG CCGTCACTCG GATCGGGGGA CCTCCCTTGG
151  GAAATCAATC CCCGTCCTCC TGTTCTTTGC TCCATGAGAA AGATCCACCT
201  ACGACCTCAG GTCCTCAGAC AGACCAGCCC AAGAAACATC TCACCAATTT
                SJ                          Hinc II
251  CAAATCCGCT ACCAGGAGGG TGGCCAGAAC TCAGTGGTTG ACAGCTGACA
301  GACAGACGTG GGCTTCCATA TCGTCCGTGC CCTGGGCTCA GACCATCAGT
351  GAGAAAAAAC CTGGAGGGTC TCTCTGGGAA ACTCGTTCTT CCCCACCGAC
401  TACTGCAGGG ACCGAGGAAG CCATGAACAC TACAAGCCTT TTGGCGCCTG
451  CTGCTGAGAT AATGGCCACA CCTGGCAGCC CATCCCAGGC CAGCCCTACC
```

Figure 5.1 Schematic and sequence of the 5' region of the PLA2L fusion transcript.

A.)HERV-H encoded sequences, which end at the splice junction site (SJ) are shown as a thick black line. The Hinc II site used to delete the 5' HERV-H sequences to create pPLA2L-del is shown as Δ. Possible initiating methionine residues are indicated as M1, M2, M3, respectively. The region of the PLA2L cDNA which was expressed as a GST-fusion and used to generate rabbit polyclonal antiserum is shown as a black bar. The Kozak consensus sequence aligned to the 3 potential start codons is shown within a box. The complete map and sequence of this cDNA have been published previously (Feuchter-Murthy et al. 1993). **B**.) DNA sequence of 5' terminus of the PLA2L cDNA used in Chapter 5. Potential initiating Met residues are underlined and correspond to M1-3 in **A**. Hinc II site used in deletion constructs labeled above sequence. HERV-H-derived sequence shown in bold.

further purification of the fusion protein away from the beads by competition with reduced glutathione results in a much less concentrated fusion protein solution. For this reason, and since agarose bead-bound fusion proteins have been recently shown to be potent immunogens (Oettinger et al. 1992), the PLA2L-AF3 protein, bound to beads was used in antibody generation. Agarose beads containing purified PLA2L:GST fusion protein were repeatedly washed and were used, in conjunction with Freund's incomplete adjuvant, to directly immunize a New Zealand White rabbit.

To generate pGEX-PLA, bases 753-959 of the AF-5 cDNA were PCR amplified and cloned into the EcoRI/EcoRV sites of pGEX3T, as the forward PCR primer had a EcoRI site engineered into its 5' end. This construct, termed pGEX-PLA, was determined to be in-frame to GST by sequencing. The fusion was expressed and purified as above, but visualized by staining with ethidium bromide and photographing on a UV transilluminator. This is shown in the right panel in Figure 5.2, where 15 μl of purified 33 kDa protein, bound to GSH-agarose beads, can be seen in two lanes. The native pGEX3T vector expressing GST is seen in the first lane. The reduced concentration of fusion protein relative to PLA2L-AF3 is likely due to less optimal induction.

5.2.2 Endogenous expression of PLA2L in teratocarcinoma cells

It has been previously shown by Northern analysis that HERV-H promoted PLA2L transcripts are present in NTera2D1, the cell line from which the original cDNA was isolated, and in an independent teratocarcinoma cell line, Tera1 (Feuchter-Murthy et al. 1993). The level of PLA2L mRNA is at least 10 fold higher in Tera1 than in NTera2D1. In Tera1 and NTera2D1, there is no evidence of PLA2L being transcribed by a promoter other than the HERV-H LTR. Transcription of PLA2L was not detected

in other cell lines by Northern analysis or by RT-PCR. Interestingly, these results mimic what has been observed for the population of HERV-H elements in general. HERV-H elements with an LTR structure like that of the PLA2L element are highly transcribed in teratocarcinoma cell lines and, of the cell lines tested, Tera1 has the highest level of HERV-H mRNA (Wilkinson et al. 1990).

To examine the potential translation of the PLA2L mRNA, the clarified rabbit anti-PLA2L antiserum, generated from the above PLA2L-AF3 fusion protein, was used in western blotting experiments, at a dilution of approximately 1:750. Prior to the primary boost of PLA2L:GST protein, preimmune sera was taken, and seen to be negative for anti-PLA2L reactivity. Mammalian cell lysates and western blotting were performed as previously described (Chapter 2, (Liu et al. 1994)). Translation was initially examined by attempting to detect PLA2L protein in human teratocarcinoma cell lines. Both NTera2D1 and Tera1 were assayed for the presence of PLA2L protein synthesis. However, despite the high level of PLA2L RNA (Feuchter-Murthy et al. 1993), no evidence of specific immunoreactive PLA2L protein was seen on western blots of lysates of Tera1 (Fig. 5.3) and NTera2D1 (data not shown), or immunoprecipitations of Tera1 lysates with rabbit polyclonal antiserum (data not shown). Thus, it appears that PLA2L is not translated, or translated at a much lower level than mRNA abundance would suggest.

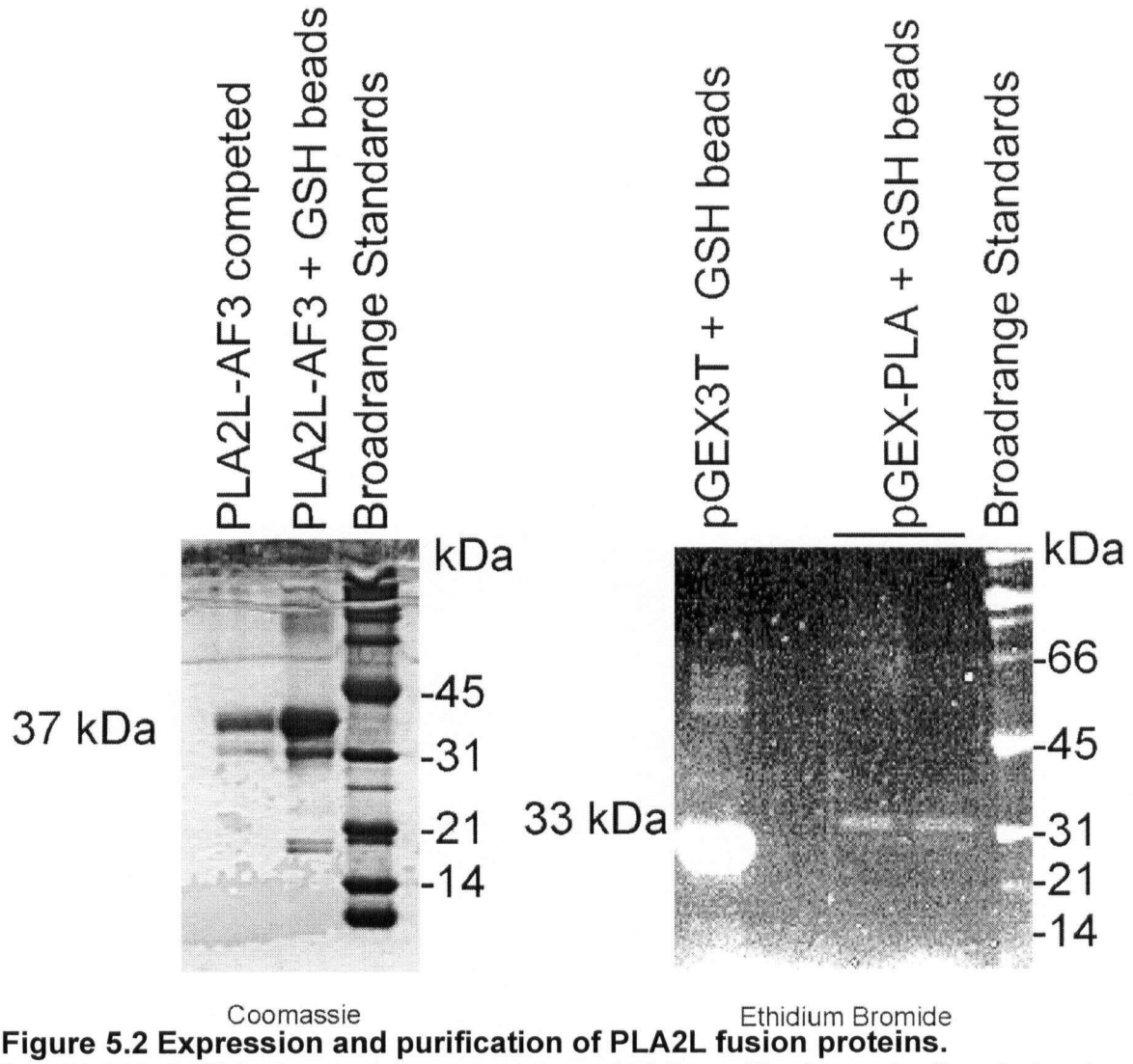


Figure 5.2 Expression and purification of PLA2L fusion proteins.
Two PLA2L:GST fusion proteins were generated for antibody production, but only PLA2L-AF3 antiserum was subsequently used. The left panel shows the purified 37 kDa PLA2L-AF3 protein, competed away from the glutathione (GSH)-agarose beads in the first lane, and bound to beads in the second lane. This gel was stained with Coomassie Blue. The right panel shows the 33 kDa pGEX-PLA fusion protein; this gel was stained with ethidium bromide. The first lane shows GST produced from empty parental vector, pGEX3T. Following a blank lane, the next two lanes are two isolates of purified pGEX-PLA protein, bound to GSH-agarose beads. Bio-Rad broad-range protein size standards were used in both gels.

5.2.3 HERV-H sequences affect translation of the PLA2L mRNA.

The lack of detectable PLA2L protein in teratocarcinoma cells which express the PLA2L mRNA raises the possibility that the presence of HERV-H sequences in the 5' UTR might inhibit translation. To test this possibility, translation of different PLA2L cDNA constructs was examined after transfection into COS cells. The mammalian expression vector pCDNA3 (Invitrogen) containing the cytomegalovirus early promoter/enhancer was used as a backbone for two PLA2L constructs. To construct pPLA2L-Full, the complete 2.4 kb AF-5 cDNA containing HERV-H sequences at the 5' end (Fig. 5.1) was cloned into the EcoRI site of pCDNA3. As our laboratory has previously noticed occasional recombination and instability of HERV-H sequences within the DH5α *E. coli* strain, this ligation was transformed into XL2-Blue (Clontech) and STBL2 (Life Technologies), two mutant *E.coli* strains known to suppress some recombinations. The constructs derived from XL2-Blue and STBL2 bacteria were termed pPLA2L-Full1 and -Full2, respectively. A 5' deletion construct, lacking all HERV-H-derived sequences and termed pPLA2L-del, was generated by inserting the 2166 bp Hinc II fragment of the PLA2L cDNA in pBluescript into the EcoRV site of pCDNA3. The vector/insert junctions of all constructs were subsequently sequenced to confirm correct orientation relative to the CMV promoter. These constructs were transiently transfected into COS cells using DEAE-dextran (Hammarskjold et al. 1986). Transfected cells were grown for 48 hours, then lysed in Nonidet P-40 lysis buffer as previously described (Liu et al. 1994). Following centrifugation, the concentration of protein in the supernatant was determined using the Bradford assay (Bio-Rad), and approximately 10 μg of COS transfectant lysates and 20 μg of Tera1 and BaF3 lysates (an irrelevant murine cell line) were electrophoresed on SDS-PAGE, electroblotted

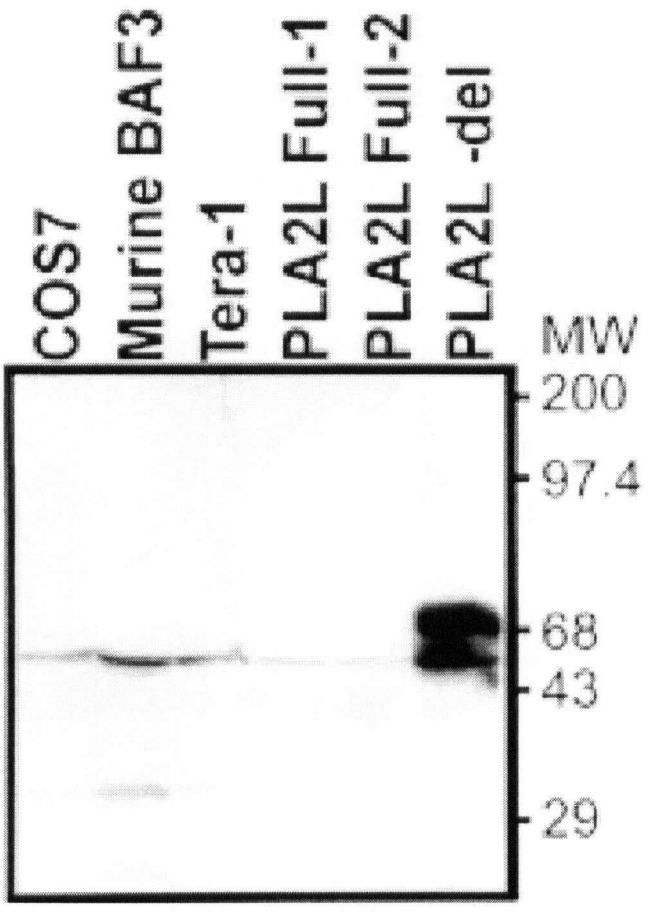


Figure 5.3 Anti-PLA2L western blot of teratocarcinoma lines and PLA2L transfectants.
The western blot was hybridized with a 1:750 dilution of PLA2L polyclonal antiserum and visualized with enhanced chemiluminescence. A positive signal is seen only in the PLA2L-del transfectant. Faint bands seen in COS7 and BaF3 are background cross-reactivity, PLA2L is not expressed in these lines as assayed by RT-PCR. Multiple bands seen in the PLA2L-del lane likely reflects usage of alternative start codons or variable post-translational modifications, and not proteolytic degradation, as they have been seen in multiple experiments. Molecular weights, in thousands of kilodaltons, are noted.

onto a PVDF membrane and western blotting was performed with anti-PLA2L antiserum. Figure 5.3 shows immunoreactive products between 65-85 kDa in the pPLA2L-del transfectant, and the lack of specific immunoreactive bands in both the pPLA2L-Full transfectants, and in the Tera1 human teratocarcinoma cell line lysates. The faint 43 kDa band seen in the COS7, BAF3 and Tera1 cell lysates is nonspecific and seen in all lysates tested, including ones negative for PLA2L transcription as assayed by RT-PCR. In addition, no specific immunoreactivity was detected in a negative control murine hemopoietic cell line, BaF3, or in the transfectant host cell line, COS. The presence of 2-3 immunoreactive bands in the pPLA2L-del transfectant likely signifies either the use of alternative AUGs to initiate translation, or differential glycosylation/modification of the PLA2L protein by COS cell systems. Random proteolytic degradation does not seem to be the cause, as similar patterns of bands are seen in multiple, independent transfections (data not shown).

5.2.4 HERV-H sequences suppress PLA2L translation, not transcription.

To ensure that the observed inhibition of PLA2L protein synthesis was due to HERV-H sequences inhibiting translation and not transcription, total RNA was prepared from all transfectants with Trizol (Life Technologies). RNA formaldehyde gel electrophoresis and northern blotting was carried out as previously described (Krosl et al. 1995) with a 410 bp BbsI fragment of the PLA2L cDNA containing the first PLA-domain (Feuchter-Murthy et al. 1993) (Fig. 5.4). Intact, full length PLA2L mRNA was seen only in the PLA2L transfectant RNAs, and not in the vector control. Figure 5.4 shows that there is a slight increase (~2 fold) in the amount of the pPLA2L-del mRNA relative to the level of pPLA2L-Full1 mRNA. This implies either that the deletion of the

HERV-H 5' UTR modestly increases the heterologous transcription (or mRNA stability) of PLA2L, or that transfection and replication of the pPLA2L-del construct was slightly more efficient. However, this modest increase cannot account for absence of detectable PLA2L protein in the pPLA2L-Full constructs. It is more probable that the deletion of HERV-H sequences in pPLA2L-del enables efficient translation of PLA2L protein.

5.2.5 HERV-H sequences do not suppress translation of a heterologous gene, Thy-1

As HERV-H sequences had been shown to specifically suppress translation of the associated PLA2L fusion transcript, the complementary experiment was to determine whether the HERV-H fragment, acting as a 5' UTR, could suppress the translation of an unrelated gene. The human hemopoietic protein Thy-1 (CD90) was chosen as a fusion partner for the suppressor 251 bp HERV-H fragment by virtue of its cell surface expression, enabling simple assaying of translation levels by fluorescence-activated cell scanning (FACS). Thy-1 is normally expressed in hemopoietic cells as a 23-35 kDa cell surface protein, specifically in the CD34+ progenitor fraction, and is not expressed in COS cells, where the HERV-H/Thy-1 chimeras were expressed (Craig et al. 1993). As detailed in the Materials and Methods chapter, a 289 bp fragment containing the suppressor HERV-H sequences from PLA2L was "swapped" for the 5' UTR of Thy-1 cDNA. As positive and negative controls, the complete Thy-1 cDNA possessing its endogenous 5' UTR, and the empty expression vector were used, respectively. In the initial experiments, the expression vector pAX142 containing the human elongation factor 1α promoter/enhancer and an SV40 origin of replication was used. These constructs were transfected into COS cells with DEAE/dextran and

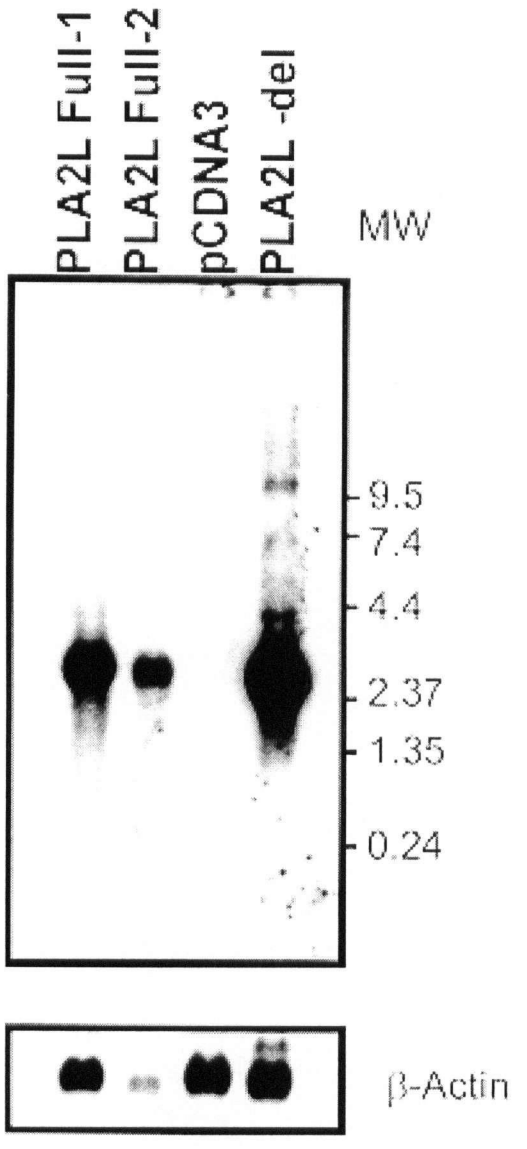


Figure 5.4 Northern blot of PLA2L transfectants.
Northern blot of total RNAs from COS cells transfected with full-length PLA2L cDNA (PLA2L-Full1 and PLA2L-Full2 respectively), vector only control (pCDNA3) and deletion construct lacking all HERV-H sequences (PLA2L-del). The lower panel shows the blot rehybridized to a β-actin probe. A 410 bp PLA-domain probe (Probe 2 in Fig. 3.1) was labeled with ^{32}P-dCTP (Amersham) and hybridized to the northern at approximately 3X10^6 dpm/mL, for 16 hours, at 42°C. After washing (detailed in Chapter 2), the filter was exposed to Kodak X-Omat AR film for 48 hours, at -70°C. A chicken β-actin cDNA probe to control for RNA loading was labeled, hybridized and washed as above, but was exposed to X-ray film for only 36 hours. RNA size ladder is shown at right, in kb.

harvested 48 hours later. The transfected cells were gently lifted and disassociated with EDTA and stained with an anti-Thy-1 monoclonal antibody (5E10) which was conjugated with the fluorochrome phycoerythrin. Following staining, the cells were loaded onto a FACScan scanner and data was acquired upon the basis of Thy-1 cell surface expression. The data was plotted onto a 2-axis dotplot, where each dot represents a discrete cell, with increasing FL2 fluorescence on the Y-axis and side scatter (SSC; a measure of cell granularity) on the X-axis. FL2 is the preset laser wavelength to acquire phycoerythrin fluorescence, and a greater proportion of dots with greater FL2 height is indicative of greater Thy-1 expression upon the transfected COS cells.

As shown in Fig. 5.5A, negative control transfection with empty pAX142 vector results in very little FL2 fluorescence. The diagonal line seen in Fig. 5.5A and all subsequent panels is a manually drawn region, to exclude cellular auto-fluorescence. All signals to the left of the region line are due to low-level COS cell autofluorescence, while almost all signals to the right of the region line represent authentic anti-Thy-1 fluorescent staining. The percentage of positive cells in the negative control experiment is 2.1%. Figure 5.5B shows the positive control transfection with normal Thy-1, and a majority of signals in the positive region (58%), between 10^1-10^4 units of FL2 fluorescence. When the 289 bp EcoRI/HincII fragment containing the PLA2L HERV-H sequences was placed just upstream of the initiating methionine of Thy-1, no effects were seen on Thy-1 translation (Fig. 5.5C). Approximately 52% of signals were positive for Thy-1 expression. These experiments, done in duplicate, proved that the HERV-H sequences suppressive of translation in PLA2L do not possess a type of ubiquitous suppressor activity. Although the translation suppression seen in PLA2L

(Fig. 5.3) is alleviated by deletion of the 5'-terminal 289 bp HERV-H fragment, the suppression cannot be transferred to a heterologous gene with the same fragment. This implies that the observed translation suppression phenomena is due to the juxtaposition of the HERV-H sequences *with* proximal unique PLA2L sequences.

As the original translation inhibition experiments were conducted using pcDNA3 as the backbone expression vector, I decided to perform an analogous set of HERV-H/Thy-1 experiments using that vector, in order to control for subtle differences or unrecognized effects upon heterologous translation with the EF1α promoter in pAX142. Vector construction specifics are found in Chapter 2, and all transfections and FACS scanning were conducted as previous. The results of the HERV-H/Thy-1 fusions using the pcDNA3 vector are seen in Figure 5.5D-F, and are analogous to the previous results. Figure 5.5D shows the FACS dotplot of empty pcDNA3 vector, while Figure 5.5E shows the positive control transfection of normal Thy-1 in pcDNA3. The proportion of signals within the positive region are 0.40% for panel D, 14% for panel E and 17% in panel F. The denser population of negative cells in panels D-F is due to less efficient transfection than in panels A-C, and to a larger number of cells analyzed. Figure 5.5F represents Thy-1 expression in COS cells transfected by pcDNA3 containing Thy-1 with the 289 bp HERV-H sequence as a 5' UTR, and as can be readily seen, there is no appreciable difference relative to normal Thy-1 expression.

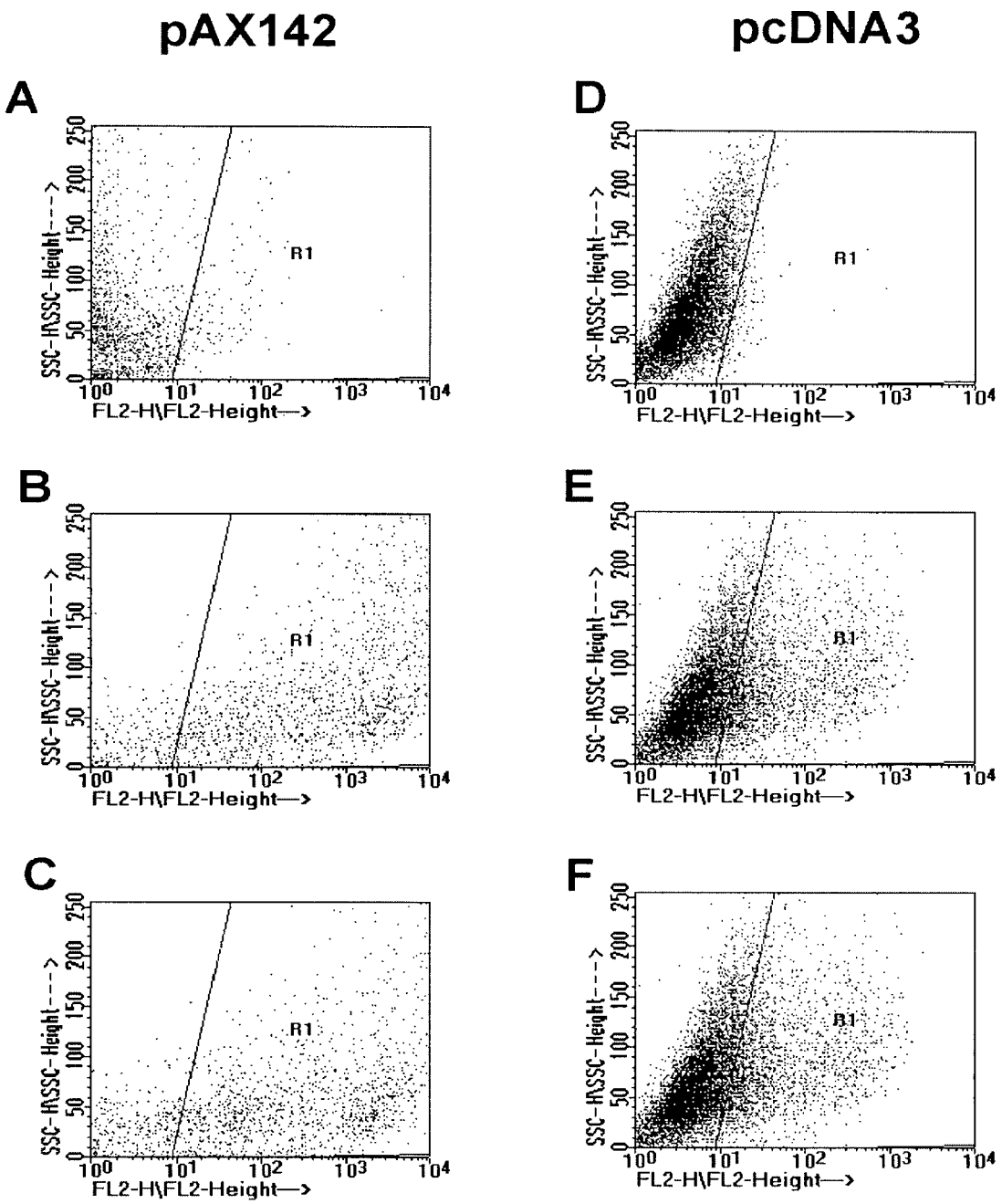


Figure 5.5 FACS dotplots of HERV-H/Thy-1 chimera transfectants.
To determine the effects of the PLA2L-derived HERV-H fragment on an unrelated gene, the 5' UTR of a reporter gene, Thy-1, was replaced with the suppressor HERV-H fragment. Effects upon translation were assayed by anti-Thy-1 FACS scanning. **(A-C)** Vector only, normal Thy-1 positive control and HERV-H/Thy-1 fusion, in pAX142, respectively. **(D-F)** As above, but using pcDNA3 as backbone. Experiments composing panels A-C are not directly comparable to those in D-F as they were performed at different times.

5.3 Discussion

The results described here led me to hypothesize that the 5' fusion of HERV-H sequences to the cellular PLA2L transcript functions as an aberrantly long and complex 5' untranslated region, explaining the concomitant inhibition of PLA2L protein synthesis. 5' UTRs are known to be the primary modulators of translation efficiency, by controlling the binding and initiation of the 43S preinitiation complex, which contains the scanning 40S ribosome, onto the 5' end of the mRNA (Pain 1996). This initiation is thought to be the rate-limiting step in translation. It has recently been shown that the 43S ribosomal complex binds mRNA by virtue of an associated protein, phospho-eIF4E, binding the ^{7}mGpppN cap structure located at the 5' end of all vertebrate mRNAs, thereby stimulating the assembly of eIF4F and the whole 43S ribosome upon the mRNA (Sachs et al. 1997). The 40S ribosome then scans the 5' UTR until an AUG codon in the correct context (Kozak 1992) is found, when the 60S subunit binds the 40S, and translation is initiated. Although the method by which the 40S ribosome scans the 5' UTR is unknown, certain structures within the 5' UTR can greatly repress or inhibit translation (Jansen et al. 1995). 5' UTRs which suppress translation generally possess some or all of the following features; stable RNA secondary structure such as stem-loops, greater length than the average of 100-140 bp, high G/C content, and AUG codons with small ORFs (μORFs) upstream of the correct start codon, especially if the μORF lacks a termination, codon (Kozak 1996).

Assuming the AUG codon with the best Kozak consensus is the most likely initiation codon for the PLA2L fusion transcript, the PLA2L fusion transcript possesses a 454 bp 5' UTR, with 252 bases being HERV-H derived. This 5' UTR contains 3 μORFs, 2 of which are in the same reading frame (+3) as the correct AUG codon, and

both of which lack a stop codon, while the third μORF exists in the +1 frame and contains a stop codon (Fig. 5.6). μORFs are hypothesized to inhibit translation by causing stalling of the scanning ribosome, while lack of a subsequent stop codon causes inefficient reinitiation of the 43S ribosome downstream at the correct AUG codon (Kozak 1992).

An additional potential encumbrance to translation which the PLA2L 5' UTR possesses is secondary RNA structures such as stem-loops. These structural elements may be the most potent translational inhibitory element found in 5' UTRs (Horvath et al. 1995; Wood et al. 1996). 5' UTRs containing secondary structures with a free energy of greater than -30 kcal/mol are known to obstruct the scanning of a mRNA by the 43S preinitiation ribosome (Kozak 1994). While prediction methods differ, the energy minimization method of Zuker (Zuker 1989) used by the RNAStructure program (Mathews et al. 1998) predicts a strong stem-loop with a free energy of -52.1 kcal/mol, in the PLA2L 5' UTR, between nt 209-369. The probability that this predicted stem-loop functions in the observed translational suppression of PLA2L is supported by the observation that the Hinc II site used to delete HERV-H sequences and construct pPLA2L-del occurs in the center of this stem-loop. Removal of sequences 5' to the site would destroy the predicted stem-loop (Fig. 5.6). The G/C nucleotide content of the PLA2L 5' UTR does not appear to differ from the average. Although deletion of the HERV-H encoded 5' terminal 251 bases releases the PLA2L transcript from translational suppression, allowing efficient heterologous protein synthesis (Fig. 5.3), unique 5' UTR sequences proximal to the junction with HERV-H also play a crucial role in translational control. Insertion of the 251 bp HERV-H

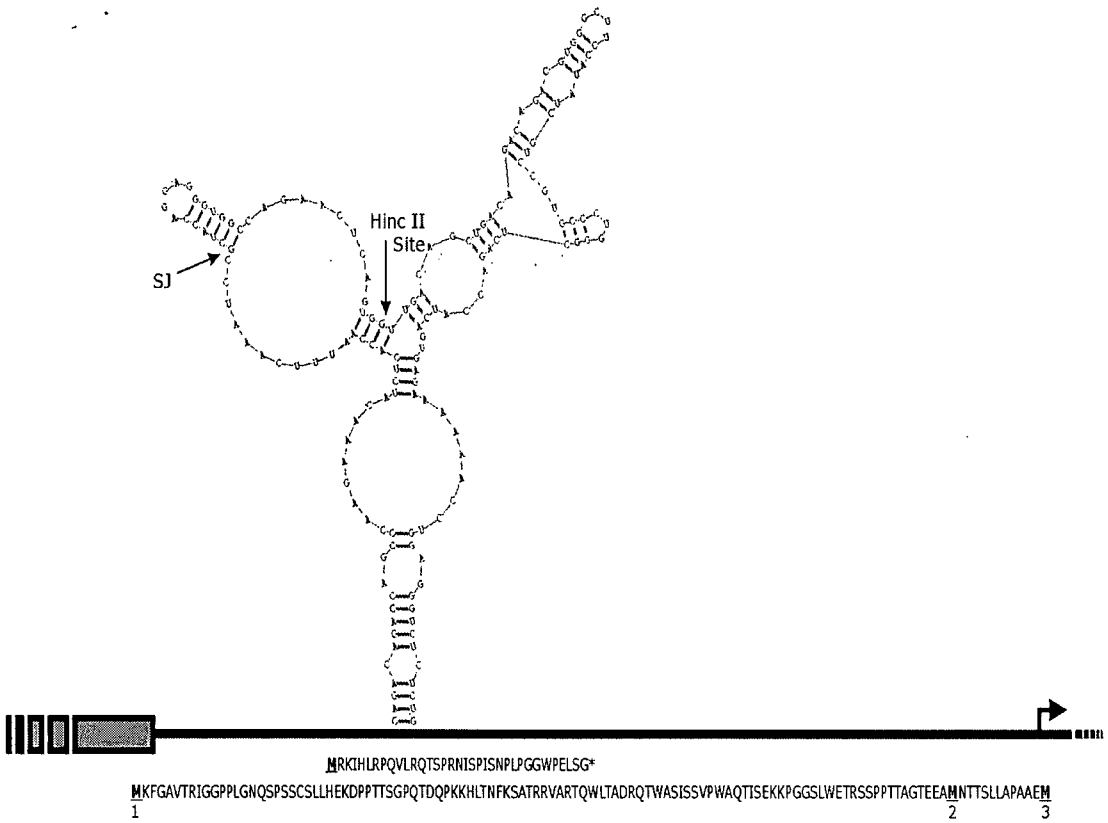


Figure 5.6 Translational inhibitory structures within the PLA2L 5'UTR.
5' region of PLA2L cDNA shown as a black line, with HERV-H LTR sequences shown as a gray box, and the junction between HERV-H and unique PLA2L sequences is denoted as SJ on the stem-loop. Small upstream open reading frames (μORFs) shown below the line. Potential initiating methionines are underlined, and numbered according to Figure 5.1. The putative start of translation is shown by an arrow. The strongest and most stable predicted RNA stem-loop is shown above the line, from nt 209-369, with a ΔG of -52.1 kcal/mol. The Hinc II site used to construct pPLA2L-del is shown on the RNA stem-loop.

fragment into the 5' UTR of a unrelated reporter gene (the human cell surface molecule Thy-1/CD90 (Craig et al. 1993)) within the same vector resulted in no change in protein expression, relative to controls (Figure 5.5). This indicates that the HERV-H fragment does not adversely affect the translation of all genes. These results suggest the juxtaposition of proximal sequences unique to the PLA2L locus and HERV-H sequences both seem necessary for the inhibition of PLA2L protein synthesis. This phenomenon is predicted by the RNA stem-loop seen in Figure 5.6, which is composed of both HERV-H and unique PLA2L sequences.

Transcriptional effects of endogenous retroviruses on cellular genes are common with numerous examples reported in mice and some in humans (Amariglio and Rechavi 1993; Wilkinson et al. 1994; Wang et al. 1997). At the PLA2L locus, it has been demonstrated previously that the HERV-H element appears to have assumed transcriptional control of the region in teratocarcinoma cells where HERV-H LTRs are highly active (Feuchter-Murthy et al. 1993). In addition, I have recent evidence suggesting that the PLA2L transcripts produced in these cells are actually HERV-H-induced fusions with two unrelated downstream genes (Chapter 4). Translational effects of retroviruses on cellular genes are much less common but a few cases have been reported. In a murine lymphoma line, it has been found that an exogenous Moloney murine leukemia retrovirus inserted into the 5' UTR of the *lck* proto-oncogene, leads to down regulation of translation (Marth et al. 1988). Similar to my PLA2L results, the suppression is removed upon deletion of the retroviral sequences from the 5' UTR. In contrast to suppression, two examples of translational activation due to 5' UTR insertion of exogenous retroviruses are known. In the first, the only other known example in a human cell line, IL-15 protein synthesis is increased in a T-

cell leukemia line due to an HTLV-1 integration in the 5' UTR of the IL-15 gene (Bamford et al. 1996). A similar event in a murine leukemia line, due to a 5' UTR murine leukemia virus insertion, results in upregulated translation of the c-Akt proto-oncogene (Wada et al. 1995).

In this study I have shown that HERV-H sequences suppress translation at the PLA2L locus. To my knowledge, this appears to be the first description of a retroviral insertion (endogenous or exogenous) repressing the translation of a human transcript. Interestingly, the transcriptional and translational effects mediated by the HERV-H element are presumably not detrimental to the species since this particular HERV-H insertion has been fixed in the primate germ line for 15-20 million years (Chapter 3). The PLA2L fusion transcript studied here has only been detected in teratocarcinoma cells where the LTR promoter is most active (Feuchter-Murthy et al. 1993). The HERV-H element appears to be within an intron so it is possible that a native promoter located 5' to the retroviral element is active in other cell types resulting in removal of the entire HERV-H element by splicing. Unfortunately, the function of the HHAG-1gene into which the HERV-H element has inserted is not known since it has no strong similarity to known genes and neither component of the PLA2L transcript is yet represented in the human EST database (as of June, 1998). However, while the functional significance of the finding reported here remains unknown at this time, it illustrates a novel way in which retroviral insertions can effect gene expression.

CHAPTER SIX: SUMMARY AND CONCLUSION

The results described in this thesis have led to a hypothesis to explain the effects of a HERV-H element upon what was originally seen as an adjacent "gene", PLA2L. Rather than the introduction of premature stop codons or polyadenylation, it appears that the primary effect of the HERV-H element upon gene expression at the PLA2L locus is to generate a teratocarcinoma-specific tripartite gene fusion, including HERV-H sequences, 6-10 exons of an alternatively spliced, conserved but anonymous gene termed HHAG-1 and the nearly complete gene encoding the human homologue of the murine inner ear structural protein otoconin-90. Otoconin-90 contains two protein domains with similarity to $sPLA_2$. Otoconin-90 and HHAG-1 are two independent transcriptional units in mouse. In human, these two genes are very likely independent on the basis of genomic cloning and murine homologue evidence and compose the PLA2L fusion transcript. This gene fusion has not occurred as a result of gross chromosomal rearrangement, but rather, is apparently mediated by a rare phenomenon known as intergenic splicing. Previously, a single *in vivo* human example, MDS1 and EVI1 has been described (Fears et al. 1996). The mechanism of intergenic splicing predicts that the 3' terminal and 5' terminal exons of both respective genes will not be part of the fusion transcript, as they lack the requisite splice donor and acceptor sites, respectively. This is illustrated in Figure 6.1, with the relevant terminal exons half-filled in orange and the resultant transcripts containing the terminal exons also orange. The structure of the tripartite transcript PLA2L (and other clones derived from the same library, see Fig. 4.9) is in accordance with this prediction: the exon which encodes what is likely the HHAG-1 3' UTR is never seen spliced to the downstream otoconin-90 exons, nor is the analogous unique 5' UTR of the murine otoconin-90 cDNA observed in the human PLA2L fusion

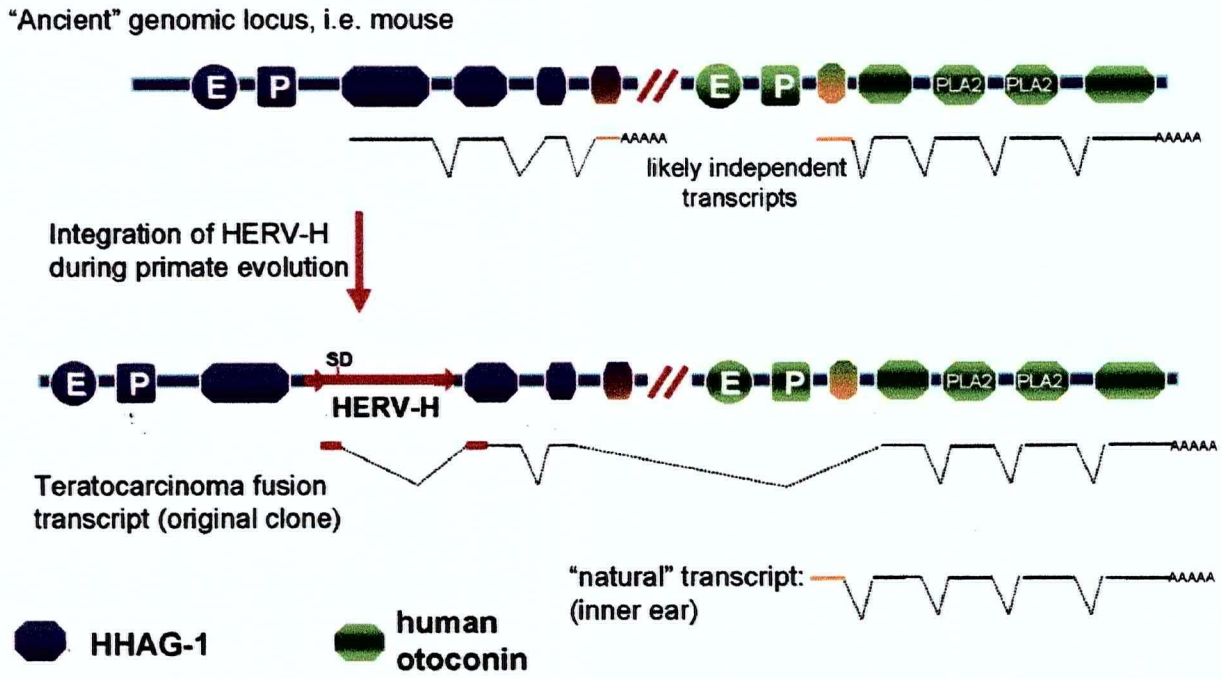


Figure 6.1 How HERV-H affects the PLA2L locus
Top panel shows the PLA2L genomic region in a species lacking HERV-H, such as mouse. The intergenic region (murine size not known, ~10 kb in human) is shown as diagonal parallel red lines. Enhancers and promoters shown as boxes containing E and P, respectively. HHAG-1 exons shown in purple, otoconin-90 in green. Genomic structure and exons shown schematically, not accurately. HHAG-1 3' terminal exon and otoconin-90 5' terminal exon are not present in the original PLA2L clone, and are distinguished by orange fill. HERV-H element in HHAG-1 intron shown in red. Translation suppressor structures shown as a thick red line in the teratocarcinoma fusion transcript.

transcript. Rather, the sequence of the fusion site between HHAG-1 and otoconin-90 in PLA2L is compatible with the predicted splicing event fusing the next-to-last exon of HHAG-1 with what is likely the second exon of human otoconin-90 (diagrammed in Fig. 4.10 and Fig.6.1).

The fusion transcript was found to be transcribed only in teratocarcinoma cell lines. This may be due to the HERV-H Type1 LTR promoter being highly active in these cells. It is unknown if promoter activity could favor intergenic splicing events. It is perhaps due to more speculative phenomena, such as suppression of normal polyadenylation leading to intergenic splicing, caused by teratocarcinoma cell lines possessing aberrant concentrations of the basal polyadenylation factors CstF-66 / CstF-77 (Colgan and Manley 1997). Provocatively, the *Drosophila* ortholog of human CstF-77 is known to enhance or suppress the effects of transposable element insertions, which are mediated by changes in polyadenylation (Takagaki and Manley 1994).

Future experiments which could address the hypothesis that PLA2L is a teratocarcinoma-specific tripartite fusion transcript caused by intergenic splicing include:

1. Test Tera1 cell line nuclear pre-mRNA using the nuclear run-off assay (Proudfoot 1989) and genomic fragments from the intergenic region (Fig. 4.10) to determine the site of HHAG-1 transcriptional termination.
2. To determine if the intergenic region is being transcribed into Tera1 pre-mRNA, test by a northern hybridization approach using intergenic genomic DNA as a probe.

3. Closely spaced genes such as HHAG-1 and otoconin-90 often possess a terminator sequence in the intergenic region to ensure efficient termination and lack of promoter interference; functionally test the PLA2L intergenic region for presence of terminators using the poly(A)-competition system of (Ashfield et al. 1994) or attempt to find a consensus binding site (G_5AG_5) for the only known human terminator, MAZ.
4. Intergenic splicing presupposes a lack of normal cleavage and polyadenylation of the upstream gene. Two poly(A) factors involved in poly(A)-signal recognition and cleavage have been shown to affect usage of alternative poly(A) signals. Examine nuclear extracts of teratocarcinoma cell lines to determine if higher or lower concentrations of the basal polyadenylation factors CstF-66 or CstF-77 exist relative to controls. Additionally, expression constructs or antisense constructs of either factor can be transfected into Tera1 cells to investigate whether altering the levels of each protein can perturb the formation of the PLA2L fusion transcript.
5. Measure functionality and efficiency of the HHAG-1 poly(A) site by incorporating it into an expression vector with a heterologous promoter and determining the resulting transcript structure in comparison to a known poly(A) signal, such as that of β-globin.
6. If RNA from human fetal inner ear otocyst can be obtained, primer extension on otoconin-90 can be performed to accurately determine the start site of transcription. Also, the unique otoconin-90 5' UTR could be isolated, using the homologous murine clone as a probe. Lacking the tissue source, it may be possible to clone the human 5' UTR / first exon from the intergenic region (its

predicted location, Fig. 4.10), on the basis of limited homology to the murine 5' UTR.

Independent of direct investigations into the mechanism of intergenic fusion via splicing, the assignment of the downstream segment of PLA2L on the basis of homology as human otoconin-90 can be functionally tested in two ways. The otoconin-specific antisera generated using the pGEX-PLA fusion protein (Chapter 5) can be used, in collaboration, for immunocytochemistry on human inner ear sections, as obtained post-mortem or by a surgical labyrinthectomy procedure to assay for specific immunoreactivity. Additionally, the predicted secretion signal in the human cDNA could be functionally tested. Constructs containing this signal at the N-terminus can be assayed for the ability to direct the secretion of epitope-tagged otoconin-90 in a heterologous COS cell system. Specific secretion versus immunoreactivity due to cell death and lysis can be examined using a specific secretion inhibitor, such as brefeldin A.

A result of this thesis was the assignment of human otoconin-90 to the distal part of the PLA2L locus. Although the goal of my thesis was to determine the role of HERV-H in the function and expression of the adjacent genes HHAG-1 and otoconin-90, much more became known about the normal expression and function of the downstream gene otoconin-90. At present, little is known about the expression of the enigmatic HHAG-1 gene. The HERV-H integration has occurred into an intron of the HHAG-1 gene and it is therefore the likeliest to suffer direct HERV-H effects, if any, upon its normal function and expression. Expression of HHAG-1 has only been seen in the LTR-directed manner in teratocarcinoma cells, with over 50 normal human tissues assayed. This is likely due to HHAG-1 possessing a narrow temporal-spatial window of

normal expression, as does otoconin-90. As many closely-clustered genes are functionally related, it is tempting to speculate that the normal function of HHAG-1 is in the developing inner ear. With over 1 kb of coding sequence and the 3' UTR known, HHAG-1 maintains its anonymity with no known homologue or EST match found to date in the Genbank databases. Further elucidation of HHAG-1's normal expression and function will have to await a greater level of depth and saturation in the contemporary EST databases.

The biological consequences of the HERV-H integration upon the HHAG-1 and otoconin-90 genes are difficult to assess, especially given the total lack of knowledge regarding the function and normal expression of the HHAG-1 gene. In teratocarcinoma cells the two genes are fused, apparently by intergenic splicing. Given the high activity of HERV-H LTR promoters in this cell type and the usage of a LTR promoter as the sole promoter of the PLA2L fusion transcript, HERV-H is implicated in the biogenesis of this fusion mRNA. Extrapolating from knowledge of the normal murine inner ear-specific expression pattern, human otoconin-90 appears mis-expressed in teratocarcinomas. The cellular and biological consequences of this mis-expression are likely negligible, as teratocarcinomas are not normal human cell types, and experiments detailed in Chapter 5 have shown that the aberrant HHAG-1/otoconin-90 fusion mRNA (PLA2L) is not detectably translated. The apparent HERV-H-mediated translational suppression of the fusion mRNA may have been selected for in evolution, in order to "neutralize" or stabilize the effects of the LTR promoter insertion. Only HERV insertions with beneficial or neutral genetic effects are conserved throughout the millions of years of primate evolution. While examination of normal human otoconin-90 expression in the inner ear epithelia was not possible, it is likely

expressed in a HERV-H-independent manner. If the HERV-H promoter/enhancer is active in the developing inner ear, the possibility exists that the aberrant fusion mRNA could be produced at some level. If this transcript escapes the translational suppression seen in teratocarcinoma and COS cell lines, it could produce a otoconin-90 fusion protein with a large N-terminal extension encoded by HHAG-1 exons. This protein would be nonfunctional due to cytoplasmic sequestration, given the lack of a secretion signal sequence at its N-terminus.

Given the high proportion of the genome devoted to HERV sequences and their apparent predilection for transcribed regions, it is important to determine their roles in human gene function and in normal human biology. These studies upon HERV-H and PLA2L have elucidated two novel phenomena. First, that a HERV-H LTR is involved in the intergenic fusion of two unrelated but adjacent genes in teratocarcinoma cells, although the LTR's role as a potential cause of this splicing is unclear. This has not been previously known to involve HERV elements. Further, the second gene was determined, using studies of the murine ortholog, to encode an inner ear protein, otoconin-90. Second, the HERV-H LTR fused to the PLA2L cDNA was shown to function to suppress the translation of the PLA2L mRNA, which is the first demonstration of a HERV having a translation-level effect upon the expression of a adjacent cellular gene. Taken together, these results contribute to and support the emerging body of work on HERV / gene interactions, and further expand the repertoire of effects to now include translation, that HERVs can have upon adjacent genes.

REFERENCES

Abrink, M., E. Larsson, and L. Hellman. 1998. Demethylation of ERV3, an endogenous retrovirus regulating the Kruppel-related zinc finger gene H-plk, in several human cell lines arrested during early monocyte development. *DNA & Cell Biology* **17**: 27-37.

Adachi, M., R. Watanabe-Fukunaga, and S. Nagata. 1993. Aberrant transcription caused by the insertion of an early transposable element in an intron of the Fas antigen gene of lpr mice. *Proceedings of the National Academy of Sciences of the United States of America* **90**: 1756-60.

Amariglio, N. and G. Rechavi. 1993. Insertional mutagenesis by transposable elements in the mammalian genome. *Environ Mol Mutagen* **21**: 212-8.

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

Ashe, M.P., L.H. Pearson, and N.J. Proudfoot. 1997. The HIV-1 5' LTR poly(A) site is inactivated by U1 snRNP interaction with the downstream major splice donor site. *EMBO Journal* **16**: 5752-63.

Ashfield, R., A.J. Patel, S.A. Bossone, H. Brown, R.D. Campbell, K.B. Marcu, and N.J. Proudfoot. 1994. MAZ-dependent termination between closely spaced human complement genes. *EMBO Journal* **13**: 5656-67.

Baban, S., J.D. Freeman, and D.L. Mager. 1996. Transcripts from a novel human KRAB zinc finger gene contain spliced Alu and endogenous retroviral segments. *Genomics* **33**: 463-72.

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

Bamford, R.N., A.P. Battiata, J.D. Burton, H. Sharma, and T.A. Waldmann. 1996. Interleukin (IL) 15/IL-T production by the adult T-cell leukemia cell line HuT-102 is associated with a human T-cell lymphotrophic virus type I region /IL-15 fusion message that lacks many upstream AUGs that normally attenuates IL-15 mRNA translation. *Proc Natl Acad Sci U S A* **93**: 2897-902.

Berget, S.M. 1995. Exon recognition in vertebrate splicing. *Journal of Biological Chemistry* **270**: 2411-4.

Birse, C.E., L. Minviellesebastia, B.A. Lee, W. Keller, and N.J. Proudfoot. 1998. Coupling termination of transcription to messenger RNA maturation in yeast. *Science* **280**: 298-301.

Britten, R.J. 1996. DNA sequence insertion and evolutionary variation in gene regulation. *Proceedings of the National Academy of Sciences of the United States of America* **93**: 9374-7.

Chen, H., I. Thalmann, J.C. Adams, K.B. Avraham, N.G. Copeland, N.A. Jenkins, D.R. Beier, D.P. Corey, R. Thalmann, and G.M. Duyk. 1995. cDNA cloning, tissue distribution, and chromosomal localization of Ocp2, a gene encoding a putative transcription-associated factor predominantly expressed in the auditory organs. *Genomics* **27**: 389-98.

Choi, Y., J.W. Kappler, and P. Marrack. 1991. A superantigen encoded in the open reading frame of the 3' long terminal repeat of mouse mammary tumour virus. *Nature* **350**: 203-7.

Church, D.M., C.J. Stotler, J.L. Rutter, J.R. Murrell, J.A. Trofatter, and A.J. Buckler. 1994. Isolation of genes from complex sources of mammalian genomic DNA using exon amplification. *Nature Genetics* **6**: 98-105.

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

Colgan, D.F. and J.L. Manley. 1997. Mechanism and regulation of mRNA polyadenylation. *Genes & Development* **11**: 2755-66.

Conrad, B., R.N. Weissmahr, J. Boni, R. Arcari, J. Schupbach, and B. Mach. 1997. A human endogenous retroviral superantigen as candidate autoimmune gene in type I diabetes. *Cell* **90**: 303-13.

Craig, W., R. Kay, R.L. Cutler, and P.M. Lansdorp. 1993. Expression of Thy-1 on human hematopoietic progenitor cells. *J Exp Med* **177**: 1331-42.

Cupillard, L., K. Koumanov, M.G. Mattei, M. Lazdunski, and G. Lambeau. 1997. Cloning, chromosomal mapping, and expression of a novel human secretory phospholipase A2. *Journal of Biological Chemistry* **272**: 15745-52.

Davidson, F.F. and E.A. Dennis. 1990. Evolutionary relationships and implications for the regulation of phospholipase A2 from snake venom to human secreted forms. *Journal of Molecular Evolution* **31**: 228-38.

Davis, J.G., J.C. Oberholtzer, F.R. Burns, and M.I. Greene. 1995. Molecular cloning and characterization of an inner ear-specific structural protein. *Science* **267**: 1031-4.

de Jong, W.W., J.A. Leunissen, and C.E. Voorter. 1993. Evolution of the alpha-crystallin/small heat-shock protein family. *Molecular Biology & Evolution* **10**: 103-26.

de Parseval, N. and T. Heidmann. 1998. Physiological knockout of the envelope gene of the single-copy ERV-3 human endogenous retrovirus in a fraction of the Caucasian population. *Journal of Virology* **72**: 3442-5.

Dennis, E.A. 1994. Diversity of group types, regulation, and function of phospholipase A2. *J Biol Chem* **269**: 13057-60.

Devereux, J., P. Haeberli, and O. Smithies. 1984. A comprehensive set of sequence analysis programs for the VAX. *Nucleic Acids Research* **12**: 387-95.

Di Cristofano, A., M. Strazullo, L. Longo, and G. La Mantia. 1995. Characterization and genomic mapping of the ZNF80 locus: expression of this zinc-finger gene is driven by a solitary LTR of ERV9 endogenous retroviral family. *Nucleic Acids Res* **23**: 2823-30.

Drwinga, H.L., L.H. Toji, C.H. Kim, A.E. Greene, and R.A. Mulivor. 1993. NIGMS human/rodent somatic cell hybrid mapping panels 1 and 2. *Genomics* **16**: 311-4.

Eickbush, T.H. 1997. Telomerase and retrotransposons: which came first? [comment]. *Science* **277**: 911-2.

Erkman, L., R.J. McEvilly, L. Luo, A.K. Ryan, F. Hooshmand, S.M. O'Connell, E.M. Keithley, D.H. Rapaport, A.F. Ryan, and M.G. Rosenfeld. 1996. Role of transcription factors Brn-3.1 and Brn-3.2 in auditory and visual system development. *Nature* **381**: 603-6.

Fan, H. 1994. Retroviruses and their role in cancer. In *The Retroviridae* (ed. J.A. Levy), pp. 313-362. Plenum Press, New York, NY.

Fears, S., C. Mathieu, N. Zeleznik-Le, S. Huang, J.D. Rowley, and G. Nucifora. 1996. Intergenic splicing of MDS1 and EVI1 occurs in normal tissues as well as in myeloid leukemia and produces a new member of the PR domain family. *Proceedings of the National Academy of Sciences of the United States of America* **93**: 1642-7.

Feng, Q., J.V. Moran, H.H. Kazazian, Jr., and J.D. Boeke. 1996. Human L1 retrotransposon encodes a conserved endonuclease required for retrotransposition. *Cell* **87**: 905-16.

Feuchter, A. and D. Mager. 1990. Functional heterogeneity of a large family of human LTR-like promoters and enhancers. *Nucleic Acids Research* **18**: 1261-70.

Feuchter-Murthy, A.E., J.D. Freeman, and D.L. Mager. 1993. Splicing of a human endogenous retrovirus to a novel phospholipase A2 related gene. *Nucleic Acids Research* **21**: 135-43.

Fogh, J. and G. Trempe. 1975. New human tumor cell lines. In *Human tumor cells in vitro* (ed. J. Fogh), pp. 115-159. Plenum Press, New York, NY.

Friedmann, I. 1976. The Mammalian Ear (ed. J.J. Head). Oxford University Press, Oxford, UK.

Furter, C.S., C.W. Heizmann, and M.W. Berchtold. 1989. Isolation and analysis of a rat genomic clone containing a long terminal repeat with high similarity to the oncomodulin mRNA leader sequence. *Journal of Biological Chemistry* **264**: 18276-9.

Garfinkel, D.J. 1992. Retroelements in microorganisms. In *The Retroviridae* (ed. J.A. Levy), pp. 107-158. Plenum Press, New York, NY.

Garson, J.A., P.W. Tuke, P. Giraud, G. Paranhos-Baccala, and H. Perron. 1998. Detection of virion-associated MSRV-RNA in serum of patients with multiple sclerosis [letter]. *Lancet* **351**: 33.

Golovkina, T.V., A. Chervonsky, J.P. Dudley, and S.R. Ross. 1992. Transgenic mouse mammary tumor virus superantigen expression prevents viral infection. *Cell* **69**: 637-45.

Golovkina, T.V., I. Piazzon, I. Nepomnaschy, V. Buggiano, M. de Olano Vela, and S.R. Ross. 1997. Generation of a tumorigenic milk-borne mouse mammary tumor virus by recombination between endogenous and exogenous viruses. *Journal of Virology* **71**: 3895-903.

Goodchild, N.L., D.A. Wilkinson, and D.L. Mager. 1992. A human endogenous long terminal repeat provides a polyadenylation signal to a novel, alternatively spliced transcript in normal placenta. *Gene* **121**: 287-94.

Goodchild, N.L., D.A. Wilkinson, and D.L. Mager. 1993. Recent evolutionary expansion of a subfamily of RTVL-H human endogenous retrovirus-like elements. *Virology* **196**: 778-88.

Gunderson, S.I., M. Polycarpouschwarz, and I.W. Mattaj. 1998. U1 snRNP inhibits pre-mRNA polyadenylation through a direct interaction between U1 70k and poly(a) polymerase. *Molecular Cell* **1**: 255-264.

Hammarskjold, M.L., S.C. Wang, and G. Klein. 1986. High-level expression of the Epstein-Barr virus EBNA1 protein in CV1 cells and human lymphoid cells using a SV40 late replacement vector. *Gene* **43**: 41-50.

Henikoff, S., E.A. Greene, S. Pietrokovski, P. Bork, T.K. Attwood, and L. Hood. 1997. Gene families: the taxonomy of protein paralogs and chimeras. *Science* **278**: 609-14.

Hirose, Y., M. Takamatsu, and F. Harada. 1993. Presence of env genes in members of the RTVL-H family of human endogenous retrovirus-like elements. *Virology* **192**: 52-61.

Hohjoh, H. and M.F. Singer. 1997. Sequence-specific single-strand RNA binding protein encoded by the human LINE-1 retrotransposon. *EMBO Journal* **16**: 6034-43.

Horvath, P., A. Suganuma, M. Inaba, Y.B. Pan, and K.C. Gupta. 1995. Multiple elements in the 5' untranslated region down-regulate c-sis messenger RNA translation. *Cell Growth & Differentiation* **6**: 1103-10.

Huber, B.T., U. Beutner, and M. Subramanyam. 1994. The role of superantigens in the immunobiology of retroviruses. *Ciba Foundation Symposium* **187**: 132-40; discussion 140-3.

Jansen, M., C.H. de Moor, J.S. Sussenbach, and J.L. van den Brande. 1995. Translational control of gene expression. *Pediatric Research* **37**: 681-6.

Johnson, L.K., S. Frank, P. Vadas, W. Pruzanski, A.J. Lusis, and J.J. Seilhamer. 1990. Localization and evolution of two human phospholipase A2 genes and two related genetic elements. In *Phospholipase A2* (ed. P.Y.-K.W.a.E.A. Dennis). Plenum Press, New York, NY.

Kambhu, S., P. Falldorf, and J.S. Lee. 1990. Endogenous retroviral long terminal repeats within the HLA-DQ locus. *Proceedings of the National Academy of Sciences of the United States of America* **87**: 4927-31.

Kato, N., K. Shimotohno, D. VanLeeuwen, and M. Cohen. 1990. Human proviral mRNAs down regulated in choriocarcinoma encode a zinc finger protein related to Kruppel. *Molecular & Cellular Biology* **10**: 4401-5.

Kazazian, H.H. and J.V. Moran. 1998. The impact of L1 retrotransposons on the human genome. *Nature Genetics* **19**: 19-24.

Kazmierczak, B., Y. Pohnke, and J. Bullerdiek. 1996. Fusion transcripts between the HMGIC gene and RTVL-H-related sequences in mesenchymal tumors without cytogenetic aberrations. *Genomics* **38**: 223-6.

Kelley, J. 1992. Evolution of apes. In *The Cambridge encyclopedia of human evolution.* (ed. S. Jones), pp. 223-230. Cambridge University Press, Cambridge, UK.

Kipersztok, S., G.A. Osawa, L.F. Liang, W.S. Modi, and J. Dean. 1995. POM-ZP3, a bipartite transcript derived from human ZP3 and a POM121 homologue. *Genomics* **25**: 354-9.

Koch, J., S. Gartner, C.M. Li, L.E. Quintern, K. Bernardo, O. Levran, D. Schnabel, R.J. Desnick, E.H. Schuchman, and K. Sandhoff. 1996. Molecular cloning and characterization of a full-length complementary DNA encoding human acid ceramidase. Identification of the first molecular lesion causing Farber disease. *Journal of Biological Chemistry* **271**: 33110-5.

Komada, M., I. Kudo, and K. Inoue. 1990. Structure of gene coding for rat group II phospholipase A2. *Biochemical & Biophysical Research Communications* **168**: 1059-65.

Kozak, M. 1992. Regulation of translation in eukaryotic systems. *Annu Rev Cell Biol* **8**: 197-225.

Kozak, M. 1994. Features in the 5' non-coding sequences of rabbit alpha and beta-globin mRNAs that affect translational efficiency. *Journal of Molecular Biology* **235**: 95-110.

Kozak, M. 1996. Interpreting cDNA sequences: some insights from studies on translation. *Mammalian Genome* **7**: 563-74.

Krosl, J., J.E. Damen, G. Krystal, and R.K. Humphries. 1995. Erythropoietin and interleukin-3 induce distinct events in erythropoietin receptor-expressing BA/F3 cells. *Blood* **85**: 50-6.

Kumar, S. and S.B. Hedges. 1998. A molecular timescale for vertebrate evolution. *Nature* **392**: 917-920.

Labuda, D., E. Zietkiewicz, and G.A. Mitchell. 1995. Alu elements as a source of genomic variation:deleterious effects and evolutionary novelties. In *The impact of short interspersed elements (SINEs) on the host genome* (ed. R.J. Maraia), pp. 1-24. R.G. Landes Co., Austin, TX.

Lania, L., A. Di Cristofano, M. Strazzullo, G. Pengue, B. Majello, and G. La Mantia. 1992. Structural and functional organization of the human endogenous retroviral ERV9 sequences. *Virology* **191**: 464-8.

Lehrman, M.A., W.J. Schneider, T.C. Sudhof, M.S. Brown, J.L. Goldstein, and D.W. Russell. 1985. Mutation in LDL receptor: Alu-Alu recombination deletes exons encoding transmembrane and cytoplasmic domains. *Science* **227**: 140-6.

Li, W.-H. and D. Graur. 1991. Fundamentals of molecular evolution. Sinauer Associates, Sunderland, MA.

Li, Y., B.Z. Yu, H. Zhu, M.K. Jain, and M.D. Tsai. 1994. Phospholipase A2 engineering. Structural and functional roles of the highly conserved active site residue aspartate-49. *Biochemistry* **33**: 14714-22.

Lindahl, K.F. 1991. His and hers recombinational hotspots. *Trends in Genetics* **7**: 273-6.

Lindeskog, M., P. Medstrand, A.A. Cunningham, and J. Blomberg. 1998. Coamplification and dispersion of adjacent human endogenous retroviral Herv-H and Herv-E elements - Presence of spliced hybrid transcripts in normal leukocytes. *Virology* **244**: 219-229.

Liu, A.Y. and B.A. Abraham. 1991. Subtractive cloning of a hybrid human endogenous retrovirus and calbindin gene in the prostate cell line PC3. *Cancer Research* **51**: 4107-10.

Liu, L., J.E. Damen, R.L. Cutler, and G. Krystal. 1994. Multiple cytokines stimulate the binding of a common 145-kilodalton protein. *Molecular & Cellular Biology* **14**: 6926-35.

Lower, R., J. Lower, and R. Kurth. 1996. The viruses in all of us: characteristics and biological significance of human endogenous retrovirus sequences. *Proc Natl Acad Sci U S A* **93**: 5177-84.

Luciw, P.A. and N.J. Leung. 1992. Mechanisms of retrovirus replication. In *The Retroviridae* (ed. J.A. Levy), pp. 159-298. Plenum Press, New York, NY.

Lyon, M.F., S. Rastan, and S.D.M. Brown. 1996. Genetic variants and strains of the laboratory mouse. Oxford University Press, Oxford, UK.

Maeda, N. and H.S. Kim. 1990. Three independent insertions of retrovirus-like sequences in the haptoglobin gene cluster of primates. *Genomics* **8**: 671-83.

Mager, D.L. 1989. Polyadenylation function and sequence variability of the long terminal repeats of the human endogenous retrovirus-like family RTVL-H. *Virology* **173**: 591-9.

Mager, D.L. and J.D. Freeman. 1987. Human endogenous retroviruslike genome with type C pol sequences and gag sequences related to human T-cell lymphotropic viruses. *Journal of Virology* **61**: 4060-6.

Mager, D.L. and J.D. Freeman. 1995. HERV-H endogenous retroviruses: presence in the New World branch but amplification in the Old World primate lineage. *Virology* **213**: 395-404.

Mager, D.L. and P.S. Henthorn. 1984. Identification of a retrovirus-like repetitive element in human DNA. *Proceedings of the National Academy of Sciences of the United States of America* **81**: 7510-4.

Maraganore, J.M. and R.L. Heinrikson. 1986. The lysine-49 phospholipase A2 from the venom of Agkistrodon piscivorus piscivorus. Relation of structure and function to other phospholipases A2 [published erratum appears in J Biol Chem 1993 Mar 15;268(8):6064]. *Journal of Biological Chemistry* **261**: 4797-804.

Marra, M.A., L. Hillier, and R.H. Waterston. 1998. Expressed sequence tags--ESTablishing bridges between genomes. *Trends in Genetics* **14**: 4-7.

Marth, J.D., R.W. Overell, K.E. Meier, E.G. Krebs, and R.M. Perlmutter. 1988. Translational activation of the lck proto-oncogene. *Nature* **332**: 171-3.

Martin, R. 1992. Classification of primates. In *The Cambridge encyclopedia of human evolution.* (ed. S. Jones), pp. 20-23. Cambridge University Press, Cambridge, UK.

Martinelli, S.C. and S.P. Goff. 1990. Rapid reversion of a deletion mutation in Moloney murine leukemia virus by recombination with a closely related endogenous provirus. *Virology* **174**: 135-44.

Mathews, D.H., T.C. Andre, J. Kim, D.H. Turner, and M. Zuker. 1998. An updated recursive algorithm for RNA secondary structure: Prediction with improved free energy parameters. In *Molecular Modelling of Nucleic Acids* (ed. N.B. Leontis and J. SantaLucia), pp. 246-257. American Chemical Society, Washington, D.C.

Mayer, W.E., C. O'HUigin, and J. Klein. 1993. Resolution of the HLA-DRB6 puzzle: a case of grafting a de novo-generated exon on an existing gene. *Proceedings of the National Academy of Sciences of the United States of America* **90**: 10720-4.

Medstrand, P. 1996. Human endogenous retroviruses: studies on transcriptional activity and genetic variability. Thesis: Dept. of Medical Microbiology. Lund, Sweden. Lund University. pp. 93.

Medstrand, P., D.L. Mager, H. Yin, U. Dietrich, and J. Blomberg. 1997. Structure and genomic organization of a novel human endogenous retrovirus family: HERV-K (HML-6). *Journal of General Virology* **78**: 1731-44.

Miki, Y., I. Nishisho, A. Horii, Y. Miyoshi, J. Utsunomiya, K.W. Kinzler, B. Vogelstein, and Y. Nakamura. 1992. Disruption of the APC gene by a retrotransposal insertion of L1 sequence in a colon cancer. *Cancer Research* **52**: 643-5.

Moran, J.V., S.E. Holmes, T.P. Naas, R.J. DeBerardinis, J.D. Boeke, and H.H. Kazazian, Jr. 1996. High frequency retrotransposition in cultured mammalian cells. *Cell* **87**: 917-27.

Moreira, A., M. Wollerton, J. Monks, and N.J. Proudfoot. 1995. Upstream sequence elements enhance poly(A) site efficiency of the C2 complement gene and are phylogenetically conserved. *EMBO Journal* **14**: 3809-19.

Mukherjee, A.B., L. Miele, and N. Pattabiraman. 1994. Phospholipase A2 enzymes: regulation and physiological role. *Biochemical Pharmacology* **48**: 1-10.

Muratani, K., T. Hada, Y. Yamamoto, T. Kaneko, Y. Shigeto, T. Ohue, J. Furuyama, and K. Higashino. 1991. Inactivation of the cholinesterase gene by Alu insertion: possible mechanism for human gene transposition. *Proceedings of the National Academy of Sciences of the United States of America* **88**: 11315-9.

Nakashima, K., T. Ogawa, N. Oda, M. Hattori, Y. Sakaki, H. Kihara, and M. Ohno. 1993. Accelerated evolution of Trimeresurus flavoviridis venom gland phospholipase A2 isozymes. *Proceedings of the National Academy of Sciences of the United States of America* **90**: 5964-8.

Nelson, D.T., N.L. Goodchild, and D.L. Mager. 1996. Gain of Sp1 sites and loss of repressor sequences associated with a young, transcriptionally active subset of HERV-H endogenous long terminal repeats. *Virology* **220**: 213-8.

Nielsen, H., J. Engelbrecht, S. Brunak, and G. von Heijne. 1997. Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. *Protein Engineering* **10**: 1-6.

Nucifora, G. 1997. The EVI1 gene in myeloid leukemia. *Leukemia* **11**: 2022-31.

Nucifora, G., C.R. Begy, H. Kobayashi, D. Roulston, D. Claxton, J. Pedersen-Bjergaard, E. Parganas, J.N. Ihle, and J.D. Rowley. 1994. Consistent intergenic splicing and production of multiple transcripts between AML1 at 21q22 and unrelated genes at 3q26 in (3;21)(q26;q22) translocations. *Proceedings of the National Academy of Sciences of the United States of America* **91**: 4004-8.

Nusse, R. 1991. Insertional mutagenesis in mouse mammary tumorigenesis. *Current Topics in Microbiology & Immunology* **171**: 43-65.

Oettinger, H.F., R. Pasqualini, and M. Bernfield. 1992. Recombinant peptides as immunogens: a comparison of protocols for antisera production using the pGEX system. *Biotechniques* **12**: 544-9.

Ohshima, K., M. Hamada, Y. Terai, and N. Okada. 1996. The 3' ends of tRNA-derived short interspersed repetitive elements are derived from the 3' ends of long interspersed repetitive elements. *Molecular & Cellular Biology* **16**: 3756-64.

O'Neill, R.J.W., M.J. O'Neill, and J.A.M. Graves. 1998. Undermethylation Associated With Retroelement Activation and Chromosome Remodelling in an Interspecific Mammalian Hybrid. *Nature* **393**: 68-72.

Pain, V.M. 1996. Initiation of protein synthesis in eukaryotic cells. *European Journal of Biochemistry* **236**: 747-71.

Patience, C., G.R. Simpson, A.A. Colletta, H.M. Welch, R.A. Weiss, and M.T. Boyd. 1996. Human endogenous retrovirus expression and reverse transcriptase activity in the T47D mammary carcinoma cell line. *Journal of Virology* **70**: 2654-7.

Patience, C., Y. Takeuchi, F.L. Cosset, and R.A. Weiss. 1998. Packaging of endogenous retroviral sequences in retroviral vectors produced by murine and human packaging cells. *Journal of Virology* **72**: 2671-6.

Patience, C., D.A. Wilkinson, and R.A. Weiss. 1997. Our retroviral heritage. *Trends in Genetics* **13**: 116-20.

Perron, H., J.A. Garson, F. Bedin, F. Beseme, G. Paranhos-Baccala, F. Komurian-Pradel, F. Mallet, P.W. Tuke, C. Voisset, J.L. Blond, B. Lalande, J.M. Seigneurin, and B. Mandrand. 1997. Molecular identification of a novel retrovirus repeatedly isolated from patients with multiple sclerosis. The Collaborative Research Group on Multiple Sclerosis. *Proceedings of the National Academy of Sciences of the United States of America* **94**: 7583-8.

Perron, H., B. Lalande, B. Gratacap, A. Laurent, O. Genoulaz, C. Geny, M. Mallaret, E. Schuller, P. Stoebner, and J.M. Seigneurin. 1991. Isolation of retrovirus from patients with multiple sclerosis [letter]. *Lancet* **337**: 862-3.

Perry, W.L., N.G. Copeland, and N.A. Jenkins. 1994. The molecular basis for dominant yellow agouti coat color mutations. *Bioessays* **16**: 705-7.

Piatigorsky, J. 1990. Molecular biology: recent studies on enzyme/crystallins and alpha-crystallin gene expression. *Experimental Eye Research* **50**: 725-8.

Pogue-Geile, K.L., J.A. Gott, and J.S. Greenberger. 1998. The role of intracisternal A-type particles in the evolution of factor-independent murine hematopoietic cell lines. *Leukemia* **12**: 4-12.

Polgar, J., E.M. Magnenat, M.C. Peitsch, T.N. Wells, and K.J. Clemetson. 1996. Asp-49 is not an absolute prerequisite for the enzymic activity of low-M(r) phospholipases A2: purification, characterization and computer modelling of an enzymically active Ser-49 phospholipase A2, ecarpholin S, from the venom of Echis carinatus sochureki (saw-scaled viper). *Biochemical Journal* **319**: 961-8.

Pote, K.G., C.R.d. Hauer, H. Michel, J. Shabanowitz, D.F. Hunt, and R.H. Kretsinger. 1993. Otoconin-22, the major protein of aragonitic frog otoconia, is a homolog of phospholipase A2. *Biochemistry* **32**: 5017-24.

Pote, K.G. and M.D. Ross. 1991. Each otoconia polymorph has a protein unique to that polymorph. *Comparative Biochemistry & Physiology - B: Comparative Biochemistry* **98**: 287-95.

Proudfoot, N. 1996. Ending the message is not so simple. *Cell* **87**: 779-81.

Proudfoot, N.J. 1989. How RNA polymerase II terminates transcription in higher eukaryotes. *Trends in Biochemical Sciences* **14**: 105-10.

Renetseder, R., S. Brunie, B.W. Dijkstra, J. Drenth, and P.B. Sigler. 1985. A comparison of the crystal structures of phospholipase A2 from bovine pancreas and Crotalus atrox venom. *Journal of Biological Chemistry* **260**: 11627-34.

Ruegg, C.L., C.R. Monell, and M. Strand. 1989. Identification, using synthetic peptides, of the minimum amino acid sequence from the retroviral transmembrane protein p15E required for inhibition of lymphoproliferation and its similarity to gp21 of human T-lymphotropic virus types I and II. *Journal of Virology* **63**: 3250-6.

Sachs, A.B. and S. Buratowski. 1997. Common themes in translational and transcriptional regulation. *Trends in Biochemical Sciences* **22**: 189-92.

Sachs, A.B., P. Sarnow, and M.W. Hentze. 1997. Starting at the beginning, middle, and end: translation initiation in eukaryotes. *Cell* **89**: 831-8.

Sambrook, J., E.F. Fritsch, and T. Maniatis. 1989. Molecular cloning: a laboratory manual. Cold Spring Harbor Laboratory, Cold Spring Harbor, NY.

Sassaman, D.M., B.A. Dombroski, J.V. Moran, M.L. Kimberland, T.P. Naas, R.J. DeBerardinis, A. Gabriel, G.D. Swergold, and H.H. Kazazian, Jr. 1997. Many human L1 elements are capable of retrotransposition [see comments]. *Nature Genetics* **16**: 37-43.

Schuler, G.D., M.S. Boguski, E.A. Stewart, L.D. Stein, G. Gyapay, K. Rice, R.E. White, P. Rodriguez-Tome, A. Aggarwal, E. Bajorek, S. Bentolila, B.B. Birren, A. Butler, A.B. Castle, N. Chiannilkulchai, A. Chu, C. Clee, S. Cowles, P.J. Day, T. Dibling, N. Drouot, I. Dunham, S. Duprat, C. East, T.J. Hudson, and et al. 1996. A gene map of the human genome. *Science* **274**: 540-6.

Schulte, A.M., S. Lai, A. Kurtz, F. Czubayko, A.T. Riegel, and A. Wellstein. 1996. Human trophoblast and choriocarcinoma expression of the growth factor pleiotrophin attributable to germ-line insertion of an endogenous retrovirus. *Proc Natl Acad Sci U S A* **93**: 14759-64.

Schulte, A.M. and A. Wellstein. 1998. Structure and phylogenetic analysis of an endogenous retrovirus inserted into the human growth factor gene pleiotrophin. *Journal of Virology* **72**: 6065-72.

Scott, D.L., A. Achari, J.C. Vidal, and P.B. Sigler. 1992. Crystallographic and biochemical studies of the (inactive) Lys-49 phospholipase A2 from the venom of Agkistridon piscivorus piscivorus. *Journal of Biological Chemistry* **267**: 22645-57.

Scott, D.L., S.P. White, J.L. Browning, J.J. Rosa, M.H. Gelb, and P.B. Sigler. 1991. Structures of free and inhibited human secretory phospholipase A2 from inflammatory exudate. *Science* **254**: 1007-10.

Seilhamer, J.J., W. Pruzanski, P. Vadas, S. Plant, J.A. Miller, J. Kloss, and L.K. Johnson. 1989a. Cloning and recombinant expression of phospholipase A2 present in rheumatoid arthritic synovial fluid. *Journal of Biological Chemistry* **264**: 5335-8.

Seilhamer, J.J., T.L. Randall, L.K. Johnson, C. Heinzmann, I. Klisak, R.S. Sparkes, and A.J. Lusis. 1989b. Novel gene exon homologous to pancreatic phospholipase A2: sequence and chromosomal mapping of both human genes. *Journal of Cellular Biochemistry* **39**: 327-37.

Seilhamer, J.J., T.L. Randall, M. Yamanaka, and L.K. Johnson. 1986. Pancreatic phospholipase A2: isolation of the human gene and cDNAs from porcine pancreas and human lung. *Dna* **5**: 519-27.

Selistre de Araujo, H.S., S.P. White, and C.L. Ownby. 1996. cDNA cloning and sequence analysis of a lysine-49 phospholipase A2 myotoxin from Agkistrodon contortrix laticinctus snake venom. *Archives of Biochemistry & Biophysics* **326**: 21-30.

Seperack, P.K., J.A. Mercer, M.C. Strobel, N.G. Copeland, and N.A. Jenkins. 1995. Retroviral sequences located within an intron of the dilute gene alter dilute expression in a tissue-specific manner. *EMBO Journal* **14**: 2326-32.

Sibley, C.G. and J.E. Ahlquist. 1987. DNA hybridization evidence of hominoid phylogeny: results from an expanded data set. *Journal of Molecular Evolution* **26**: 99-121.

Sjottem, E., S. Anderssen, and T. Johansen. 1996. The promoter activity of long terminal repeats of the HERV-H family of human retrovirus-like elements is critically dependent on Sp1 family proteins interacting with a GC/GT box located immediately 3' to the TATA box. *J Virol* **70**: 188-98.

Smit, A.F. 1996. The origin of interspersed repeats in the human genome. *Current Opinion in Genetics & Development* **6**: 743-8.

Soderholm, J., H. Kobayashi, C. Mathieu, J.D. Rowley, and G. Nucifora. 1997. The leukemia-associated gene MDS1/EVI1 is a new type of GATA-binding transactivator. *Leukemia* **11**: 352-8.

Stinissen, P., J. Raus, and J. Zhang. 1997. Autoimmune pathogenesis of multiple sclerosis: role of autoreactive T lymphocytes and new immunotherapeutic strategies. *Critical Reviews in Immunology* **17**: 33-75.

Takagaki, Y. and J.L. Manley. 1994. A polyadenylation factor subunit is the human homologue of the Drosophila suppressor of forked protein. *Nature* **372**: 471-4.

Taruscio, D. and L. Manuelidis. 1991. Integration site preferences of endogenous retroviruses. *Chromosoma* **101**: 141-56.

Temin, H. 1992. Origin and general nature of retroviruses. In *The Retroviridae* (ed. J.A. Levy), pp. 1-18. Plenum Press, New York, NY.

Ting, C.N., M.P. Rosenberg, C.M. Snow, L.C. Samuelson, and M.H. Meisler. 1992. Endogenous retroviral sequences are required for tissue-specific expression of a human salivary amylase gene. *Genes Dev* **6**: 1457-65.

Tisch, R. and H. McDevitt. 1996. Insulin-dependent diabetes mellitus. *Cell* **85**: 291-297.

Tischfield, J.A. 1997. A reassessment of the low molecular weight phospholipase A2 gene family in mammals. *Journal of Biological Chemistry* **272**: 17247-50.

Tischfield, J.A., Y.R. Xia, D.M. Shih, I. Klisak, J. Chen, S.J. Engle, A.N. Siakotos, M.V. Winstead, J.J. Seilhamer, V. Allamand, G. Gyapay, and A.J. Lusis. 1996. Low-molecular-weight, calcium-dependent phospholipase A2 genes are linked and map to homologous chromosome regions in mouse and human. *Genomics* **32**: 328-33.

Tonjes, R.R., R. Lower, K. Boller, J. Denner, B. Hasenmaier, H. Kirsch, H. Konig, C. Korbmacher, C. Limbach, R. Lugert, R.C. Phelps, J. Scherer, K. Thelen, J. Lower, and R. Kurth. 1996. HERV-K: the biologically most active human endogenous retrovirus family. *Journal of Acquired Immune Deficiency Syndromes & Human Retrovirology* **13**: S261-7.

Vanin, E.F., P.S. Henthorn, D. Kioussis, F. Grosveld, and O. Smithies. 1983. Unexpected relationships between four large deletions in the human beta-globin gene cluster. *Cell* **35**: 701-9.

Venables, P.J., S.M. Brookes, D. Griffiths, R.A. Weiss, and M.T. Boyd. 1995. Abundance of an endogenous retroviral envelope protein in placental trophoblasts suggests a biological function. *Virology* **211**: 589-92.

Vetter, D.E., J.R. Mann, P. Wangemann, J. Liu, K.J. McLaughlin, F. Lesage, D.C. Marcus, M. Lazdunski, S.F. Heinemann, and J. Barhanin. 1996. Inner ear defects induced by null mutation of the isk gene. *Neuron* **17**: 1251-64.

Vidaud, D., M. Vidaud, B.R. Bahnak, V. Siguret, S. Gispert Sanchez, Y. Laurian, D. Meyer, M. Goossens, and J.M. Lavergne. 1993. Haemophilia B due to a de novo insertion of a human-specific Alu subfamily member within the coding region of the factor IX gene. *European Journal of Human Genetics* **1**: 30-6.

Villareal, L.P. 1997. On viruses, sex, and motherhood. *Journal of Virology* **71**: 859-65.

von Heijne, G. 1985. Signal sequences. The limits of variation. *Journal of Molecular Biology* **184**: 99-105.

von Heijne, G. 1986. A new method for predicting signal sequence cleavage sites. *Nucleic Acids Research* **14**: 4683-90.

Wada, H., M. Matsuo, A. Uenaka, N. Shimbara, K. Shimizu, and E. Nakayama. 1995. Rejection antigen peptides on BALB/c RL male 1 leukemia recognized by cytotoxic T lymphocytes: derivation from the normally untranslated 5' region of the c-akt proto-oncogene activated by long terminal repeat. *Cancer Research* **55**: 4780-3.

Wagner, M.J., Y. Ge, M. Siciliano, and D.E. Wells. 1991. A hybrid cell mapping panel for regional localization of probes to human chromosome 8. *Genomics* **10**: 114-25.

Wahle, E. and U. Kuhn. 1997. The mechanism of 3' cleavage and polyadenylation of eukaryotic pre-mRNA. *Progress in Nucleic Acid Research & Molecular Biology* **57**: 41-71.

Wang, X.Y., L.S. Steelman, and J.A. McCubrey. 1997. Abnormal activation of cytokine gene expression by intracisternal type A particle transposition: effects of mutations that result in autocrine growth stimulation and malignant transformation. *Cytokines Cell Mol Ther* **3**: 3-19.

Wang, Y., P.E. Kowalski, I. Thalmann, D.M. Ornitz, D.L. Mager, and R. Thalmann. submitted. Mammalian otoconin-90 encodes a phospholipase A2 homologue. .

Weber, G.F., S. Abromson-Leeman, and H. Cantor. 1995. A signaling pathway coupled to T cell receptor ligation by MMTV superantigen leading to transient activation and programmed cell death. *Immunity* **2**: 363-72.

Weiss, R.A. 1993. Cellular receptors and viral glycoproteins involved in retrovirus entry. In *The Retroviridae* (ed. J.A. Levy), pp. 1-108. Plenum Press, New York, NY.

Wheeler, A.P. 1992. Mechanisms of molluscan shell formation. In *Calcification in Biological Systems* , pp. 179-216. CRC Press, Boca Raton, FL.

Wilkinson, D.A., J.D. Freeman, N.L. Goodchild, C.A. Kelleher, and D.L. Mager. 1990. Autonomous expression of RTVL-H endogenous retroviruslike elements in human cells. *Journal of Virology* **64**: 2157-67.

Wilkinson, D.A., N.L. Goodchild, T.M. Saxton, S. Wood, and D.L. Mager. 1993. Evidence for a functional subclass of the RTVL-H family of human endogenous retrovirus-like sequences. *Journal of Virology* **67**: 2981-9.

Wilkinson, D.A., D.L. Mager, and J.C. Leong. 1994. Endogenous human retroviruses. In *The Retroviridae* (ed. J.A. Levy), pp. 465-535. Plenum Press, New York, NY.

Wood, M.W., H.M. VanDongen, and A.M. VanDongen. 1996. The 5'-untranslated region of the N-methyl-D-aspartate receptor NR2A subunit controls efficiency of translation. *Journal of Biological Chemistry* **271**: 8115-20.

Wood, S., M. Schertzer, H. Drabkin, D. Patterson, J.L. Longmire, and L.L. Deaven. 1992. Characterization of a human chromosome 8 cosmid library constructed from flow-sorted chromosomes. *Cytogenetics & Cell Genetics* **59**: 243-7.

Zehetner, G. and H. Lehrach. 1994. The Reference Library System--sharing biological material and experimental data. *Nature* **367**: 489-91.

Zheng, N. and L.M. Gierasch. 1996. Signal sequences: the same yet different. *Cell* **86**: 849-52.

Zuker, M. 1989. On finding all suboptimal foldings of an RNA molecule. *Science* **244**: 48-52.

Zupan, L.A., D.L. Steffens, C.A. Berry, M. Landt, and R.W. Gross. 1992. Cloning and expression of a human 14-3-3 protein mediating phospholipolysis. Identification of an arachidonoyl-enzyme intermediate during catalysis. *Journal of Biological Chemistry* **267**: 8707-10.