**CHARACTERIZATION OF IAPLTR1 SUBCLASSES AND**

**BIDIRECTIONAL PROMOTER ACTIVITY:**

**"MAKING *SENSE* OF IT ALL"**

by

Natasha W. Little

B.Sc.Honours, Trinity Western University, 2013

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF

THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF SCIENCE

in

THE FACULTY OF GRADUATE AND POSTDOCTORAL STUDIES

(Medical Genetics)

THE UNIVERSITY OF BRITISH COLUMBIA

(Vancouver)

April 2017

## Abstract

Endogenous retroviruses (ERVs) are over five-times more prevalent than gene coding sequences in the mouse and human genomes (5). The long terminal repeats (LTRs) of these elements are promoter-enhancers that can have many regulatory effects on the host genome (3). Intracisternal A-Particles (IAPs) are a highly active, murine-specific class of ERV that is known to have strong LTR-driven promoter activity (10). In a recent study, a sequence divergence-based subclass nomenclature was suggested for several classes of IAPLTR – including IAPLTR1 (59). However, it remained unknown whether these high (H1) and low (L1) divergence subclasses provided any more biologically relevant information than the current class system (with no subclasses). Some, but not all, IAPLTRs can initiate sense and antisense transcripts and have thus been considered bidirectional promoters; however, due to growing interest in their form and function, a set of criteria has recently been established for what defines a bidirectional promoter and bidirectional reporter constructs have been developed. The research presented in this thesis provides a detailed analysis of bidirectional and unidirectional reporter constructs and provides evidence that bidirectional reporter constructs should be used when assaying bidirectional promoters. Using a bidirectional reporter construct, IAPLTR1 was determined to meet the criteria for a bidirectional promoter. The core promoters for IAPLTR1 and putative transcription factor binding sites that were unique to each subclass were identified. Point mutagenesis experiments revealed that functional divergence accompanied the sequence divergence of H1 and L1 subclasses. In fact, 95% of the total promoter activity of an L1 LTR was removed by changing three base-pairs to resemble the same region of an H1 LTR. The reciprocal experiment resulted in the H1 LTR losing 100% of antisense promoter activity, but

maintaining 100% of sense promoter activity. These experiments, among others, provide

evidence that the subclass system provides more biologically relevant information than the single

IAPLTR1 class, and therefore should be adopted as part of IAP nomenclature.

## Preface

Liane Gagnier made the pLucRLuc constructs that contained: full-length Wnt9bLTR in both orientations, each of the three sequences of interest, and all three Wnt9bLTR point mutation constructs. Liane also made the Wnt9b+IAP DNA prep from which the Wnt9bLTR was amplified for insertion into constructs.

Irina Wasilewitsch made the pGL3 constructs that contained IL3LTR in both orientations.

The pLucRLuc vector was provided by David Riesman, University of South Carolina.

Several plasmids were ordered from Addgene (as indicated in the text).

All other elements of this research, including project and experimental design, laboratory work, and data analysis were performed by Natasha W. Little.

# Table of Contents

# List of Tables

# List of Figures

## List of Abbreviations

5'-RACE     - 5' Rapid Amplification of cDNA Ends

aXRV     - ancient exogenous retrovirus

ASP     - HIV-1 Antisense Protein

bp     - base pair

BRE     -B Recognition Element

DPE     - Downstream Promoter Element

DNA     - deoxyribonucleic acid

eRNA     - enhancer RNA

ERV     - Endogenous Retrovirus

ETn     - Early Transposon Endogenous Retroviruses

GABP     - GA-Binding Protein (also known as Nuclear Respiratory Factor 2)

H1     - high divergence subclass of IAPLTRs (from Rep-bases' consensus sequence)

HBZ     - HTLV-1 bZIP factor

HERV     - Human Endogenous Retrovirus

HHLA     - HERV-H LTR Associating Protein

HIV     - Human Immunodeficiency Virus

HTLV     - Human T-Lymphotrophic Virus

kb     - kilobase (1000bp)

KRAB     - Krüppel Associated Box

L1     - low divergence subclass of IAPLTRs (from Rep-bases' consensus sequence)

LCA     - Last Common Ancestor

LINE           - Long Interspersed Nuclear Element

lncRNA         - long non-coding RNA

LTR            - Long Terminal Repeat

MERV           - Murine Endogenous Retrovirus

MUSCLE         - Multiple Sequence Comparison by Log-Expectation

nt             - nucleotide

ORF            - open reading frame

PCR            - polymerase chain reaction

SINE           - Short Interspersed Nuclear Element

siRNA          - small interfering RNA

SOI            - Sequence of Interest

SP1            - Specificity Protein 1

TE             - Transposable Element

TF             - Transcription Factor

TFBS           - Transcription Factor Binding Site

TSS            - transcription start site

TTS            - transcription termination site

R              - Repeated or Redundant region (present at both ends of the IAP mRNA)

RNA            - ribonucleic acid

RT-PCR         - Reverse Transcriptase Polymerase Chain Reaction

U3/U5          - Unique Region 3'/Unique Region 5' (unique to either end of the IAP mRNA)

YY1            - Ying Yang 1

ZFP            - Zinc Finger Protein

# Glossary

Balb/c        - common inbred mouse line; albino; TH-2 biased inflammatory response

C3H         - common inbred mouse line; has had the most germline IAP activity documented

C57BL6     - common inbred mouse line; considered 'genetic background' for murine model

ENV         - envelope protein; forms glycoproteins for the viral envelope

To Exapt     - the use of a feature for something other than it was originally selected for

H3K4me3   - histone modification associated with active promoters

H3K9me2   - histone modification associated with repressed DNA

H3K9me3   - histone modification associated with repressed DNA

MilliDiv     - a 1bp difference between two sequences for every 1000bp

# Acknowledgements

Thank you to my advisory committee for its support and insight at every stage of this project. And special thanks to Dr. Dixie Mager who entrusted me with significant independence on this project, embraced my eternally inquisitive nature, and patiently met abstract questions with grounded responses for two and half years. I have never known a group of individuals who have been more generous with their time. Thank you all for your commitment to this project and your investment in my education.

Thank you to Dr. Katharina Rothe, who made important suggestions that improved my statistical analysis; to Artem Babaian, who tested some of my ideas *in silico* and always provided insight for data analysis; to Liane Gagnier, who was always ready to help with anything; to Dr. Rita Rebollo, who taught me about cell culture; to Dr. Frances Lock, who did trouble-shooting with me; and to all my peers in the Terry Fox Laboratory, who were always ready to provide feedback and support.

Thank you to the professors and administrators in the department of Medical Genetics who provided the space in which I could explore this exciting field and grow in knowledge and understanding.

Thank you also to the administrators, coordinators, and other employees of the Terry Fox Laboratory who work tirelessly to ensure that there is a safe and clean lab space with well-maintained equipment, that research funding is acquired and used well, and that we have many networking and development opportunities.

Thanks is also due to my family and friends who have supported me throughout my life and have been unwavering in their support for the duration of this degree.

This research, along with my life, is dedicated to the LORD.

In the pursuit of understanding and communicating Truth.

# Chapter 1: Introduction

## 1.1   Transposable Elements

While less than 2% of human and mouse genomes are comprised of genic coding exons, approximately half of these genomes are composed of transposable elements (TEs) (Figure 1.1) (1, 2). TEs are units of DNA that have or had the ability to insert themselves into new locations in the genome, earning them the nickname 'jumping genes' (3).

There are two classes of TE: DNA transposons, which can 'cut-and-paste' themselves within a host genome; and retrotransposons, which can 'copy-and-paste' themselves within a host genome by using an RNA intermediate (3). Due to their replicative ability, retrotransposons comprise about 95% of TEs in humans and mice (4, 5). This class of TE is made up of Long and Short Interspersed Nuclear Elements (LINEs and SINEs) which have no long terminal repeats (non-LTR retroelements), and Endogenous Retroviruses (ERVs) which have long terminal repeats (LTR retroelements) (3).

## 1.2   Endogenous Retroviruses

Endogenous retroviruses (ERVs) are believed to be remnants of ancient exogenous retroviral infections (6). Exogenous retroviruses became endogenous by invading the germlines of their hosts, thereby enabling vertical transmission (from parent to offspring) (3, 7). After invading the germline, ERVs usually acquire mutations that incapacitate their ability to infect other cells, thus disabling horizontal transmission (between unrelated cells or individuals) (3, 7). Akin to the proviral form of their exogenous cousins, ERV sequences typically have *gag* and *pol* genes 'bookended' by LTRs which sit at the 5' and 3' ends of the integrated (proviral) retroviral sequence (reverse transcriptase, ribonucleaseH, and integrase genes are embedded within the *pol* gene) (Figure 1.2) (4, 8). Unlike exogenous retroviruses, ERVs often lack a functional *env* gene

1

## Human Genome

LINEs, 20.42%

SINEs, 13.29%

ERVs, 8.29%

Other, 55.16%

DNA Transposons, 2.84%

## Mouse Genome

LINEs, 19.21%

SINEs, 8.22%

ERVs, 9.87%

Other, 61.82%

DNA Transposons, 0.88%
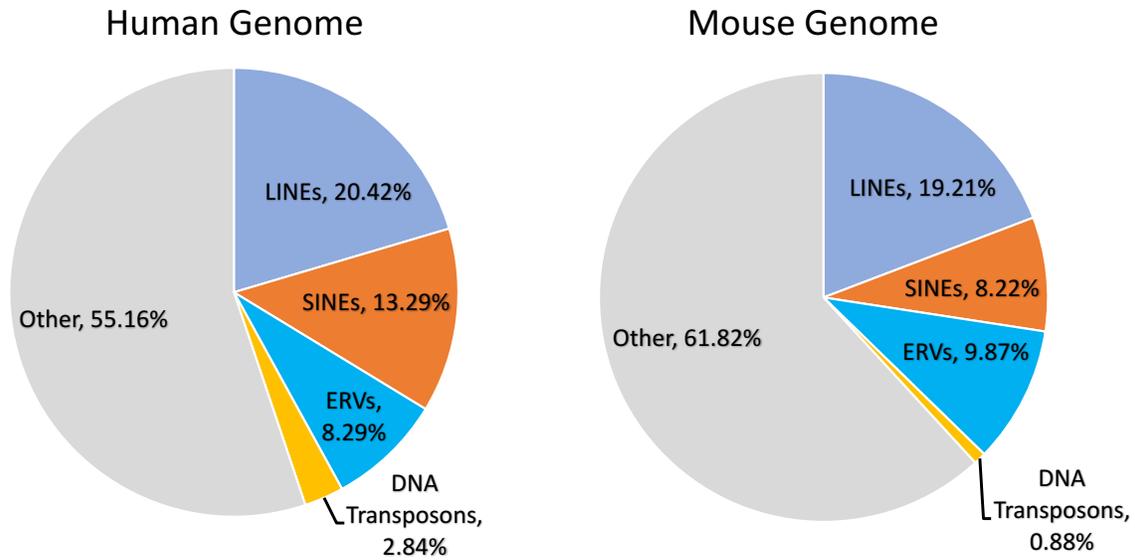
**Figure 1.1 Transposable Element Composition of the Human and Mouse Genomes**.

While less than 2% of either of these genomes is comprised of gene coding sequences, nearly half of each genome is comprised of transposable elements.

[Data from reference 5]

(7). Loss of functional *env* is typically what disables horizontal transmission of an 'endogenized' retrovirus (4, 7, 8).

Within a host genome, ERVs can be present in three main forms: as a full-length (two LTRs separated by some internal ERV sequence), as a deletion element (a full-length ERV with a deletion(s) in the internal sequence), or as a solitary LTR (Figure 1.2, Figure 1.3). Full-length, replication-competent elements replicate to colonize the host genome by: 1. Recruiting a host RNA polymerase to the 5' LTR (the retrovirus' enhancer-promoter) to transcribe the provirus into mRNA; 2. Using the virally encoded reverse transcriptase to copy the mRNA into cDNA; 3.

2

Using the encoded ribonuclease H to degrade the RNA from the DNA-RNA heteroduplex; 4. Using the virally encoded integrase to integrate the final double-stranded cDNA copy of the provirus into a new genomic location (7, 8). In contrast, solitary LTRs are created when recombination between LTRs results in excision of proviral internal sequences from the host genome, leaving a lone LTR in place of the provirus (3). 85% of ERV loci in the human genome are solitary LTRs (Figure 1.2) (1).

**1.3    ERV Effects on the Host Genome**

ERV sequences, whether a full-length sequence or a solitary LTR, can affect the host genome in a multitude of ways. One of the most popularized ways ERVs can affect a host is through insertional mutagenesis: when a new retrotransposition event causes a mutation (usually phenotypic) in the host. In mice, 10-12% of spontaneous mutations are caused by the occurrence of a new ERV insertion (10). Classic cases of insertional mutagenesis in the mouse germ line have produced kinky tails, obesity, and coat colour changes (11, 12, 13). In contrast, there is no evidence that human ERVs still retrotranspose (14, 6). The most recent germline jump of HERV-K, the youngest human ERV, was thought to be at least 100,000 years ago (6, 15, 14).

Perhaps even more captivating than cases of insertional mutagenesis are cases of protein domestication: when the host uses an ERV's protein to perform normal and, sometimes, necessary functions for the host. The Fv1 protein, for example, was one of the first retroviral resistance proteins to be identified and is derived from the gag region of an ancient MERV-L retrovirus in mice; it restricts a wide range of exogenous viruses including spumaviruses, gammaretroviruses, foamy viruses, and lentiviruses (16, 8). Fv1 is used by the host for the host's purposes, yet is derived from an ancient ERV sequence.
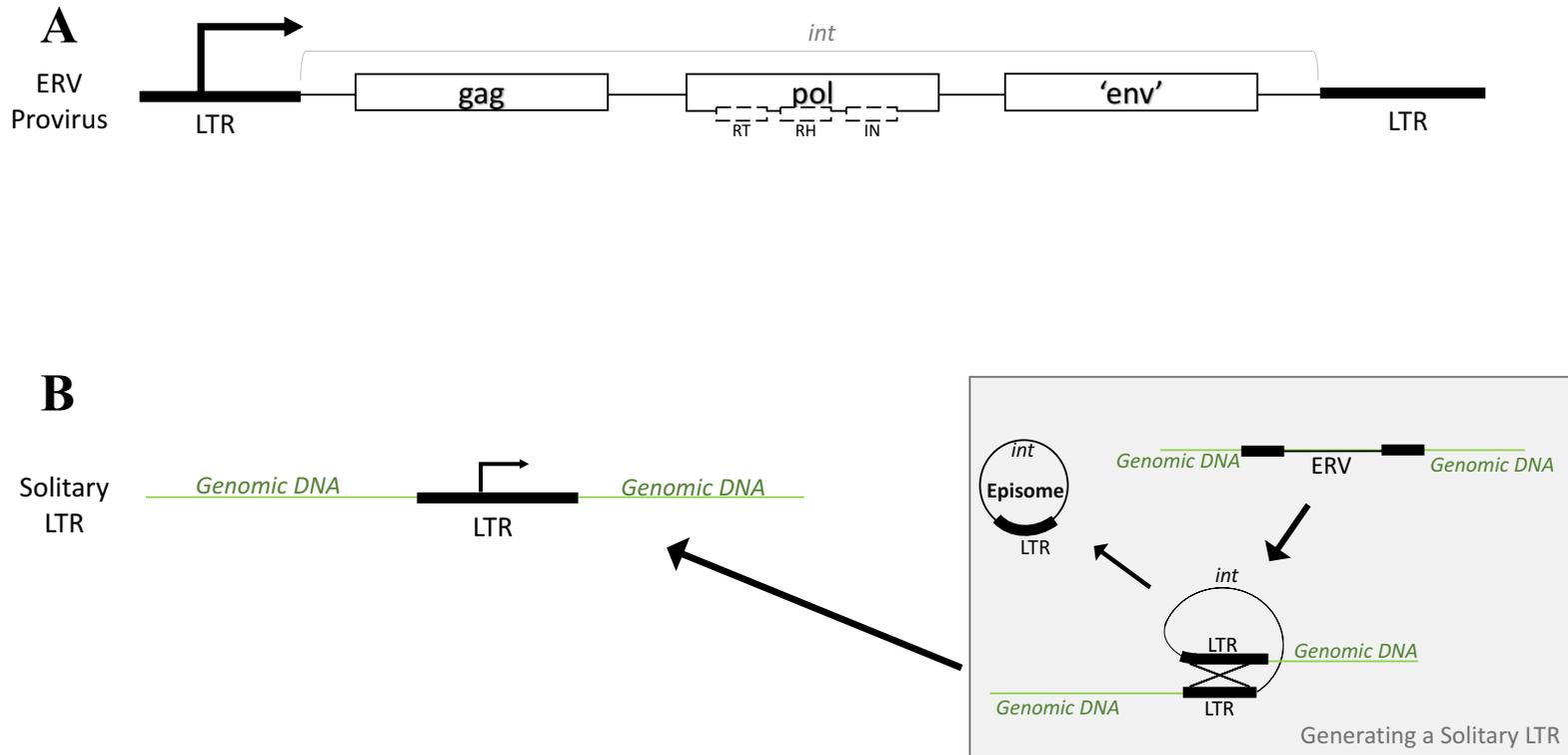
**Figure 1.2 Main ERV Forms Present in the Genome.**

**(A) A Full-Length ERV** is composed of an internal ERV sequence with an LTR on either end. The 5' LTR functions as the promoter-enhancer for transcription

of the internal ERV region (*int*). This region typically contains at least remnants of a gag gene, a pol gene (with reverse transcriptase (RT), ribonucleaseH (RH),

and integrase (IN) embedded within pol), and an incapacitated env gene (indicated by 'env'). **(B) A Solitary LTR** is generated by recombination between the 5'

and 3' LTRs. This process leaves a single LTR with no internal ERV sequence. Thick black Bars are LTRs, white boxes are coding regions, small white boxes

with dashed outlines are genes embedded within a larger coding sequence, black arrow indicates TSS, 'int' signifies the internal region of the ERV, green

indicates genomic DNA, black 'x' indicates recombination.

4

Cases of insertional mutagenesis and protein domestication are popular because of their dramatic storylines and their relatively immediate and obvious effects on the host. However, the most pervasive effects ERVs have on their hosts are the often-subtle gene regulatory effects (7, 17, 18, 19). These effects are sometimes caused by insertional mutagenesis, but in most cases of gene-regulatory effects caused by ERVs, the host utilizes ERV sequences that are already embedded in the host's genome (not newly retrotransposed elements). In fact, depending on cell-type, up to 80% of ERV sequences in the genome of a given species are in euchromatic regions (19). Perhaps unsurprising given their location in open-chromatin, ERVs have been exapted as gene-regulatory elements more frequently than any other TE (19). However, presence in open-chromatin does not completely account for the propensity of ERVs to be exapted for gene-regulation since other TEs that are less frequently exapted for gene-regulation are also located in open chromatin (1, 19).

ERVs located within introns or just down-stream of host genes can provide alternative splice sites. Splice acceptor and donor sites are often a part of the canonical ERV sequence but they can also be created through mutation over time or, rarely, during integration into the host genome (20). In humans, an ERV-K insertion into FLT4/VEGFR-3 (Fms-related Tyrosine Kinase 4/Vascular Endothelial Growth Factor Receptor 3) creates a shorter isoform of the endothelial angiogenesis controlling receptor with a different function than the longer isoform (21). A splice donor site within the LTR of the ERV is used to create a transcript variant with an alternative final exon, endowing this protein isoform with its unique functional properties (21).

These intronic and down-stream ERVs can also provide poly(A) sites for transcription termination. These sites are naturally present in the LTRs of ERVs and, when present in the same sense as the gene (poly(A) signals can only be read one way), poly(A) sites can be co-opted to

terminate transcription of a host gene. This is true in the case of human HHLA2 and HHLA3 (HERV-H LTR-Associating Protein 2 and 3) (22). These transcripts utilize the poly(A) signal present in the LTR of a HERV-H to terminate their transcript (22).

Probably due to the presence of transcription termination and splice sites in LTRs, most ERVs within and close to genes are in the antisense orientation relative to the gene. This is thought to be due to selection against the early truncation of gene products (which would result from a sense-oriented poly(A) signal too early in the gene) (3). However, antisense ERVs within introns and down-stream of genes can cause a reduction in the number of gene transcripts; this was observed with the human ERV-K insertion into IFT172 (17). The reason for this reduction in transcripts is unknown. There are two main hypotheses to consider: 1. the ERV actually causes a reduction in genic transcription rates by reducing the processivity of the polymerase (23); or 2. the antisense transcripts complimentary to regions of the gene's mRNA initiate from the antisense LTR and cause RNA interference-mediated degradation of gene transcripts (17, 24). The second hypothesis is strengthened by the fact that some ERV-initiated transcripts have known interactions with RNA interference machinery (*i.e.* DICER) and that nearly 49, 000 human gene antisense transcripts initiate in TEs (10, 17, 24).

Due to their 'ready-made' and controllable promoters (LTRs), ERVs are often exapted to function as alternative and, in some cases, primary promoters. In the case of the human aromatase gene CYP19A1, an upstream ERV-1 element provides an alternative promoter for placenta-specific expression (26, 27). The exaptation of ERV-derived promoters for tissue-specific expression is common (9). In fact, some ERV-derived promoters are so stage and cell-type specific that the repetitive transcriptome can be used to identify the stage and type of embryonic cells (28, 29). In mice, for example, a set of ERVs is used as alternative promoters to

drive transcription of the early development genes *Nfil3*, *RpI41*, and *Dnajc11* in a stage-specific manner (30). Retrotransposon-associated transcription start sites are two times more likely than other transcription start sites to have spatially or temporally restricted expression (25). This specificity alongside the strength of ERV-derived promoters is also beneficial for primary promoter activity. For example, in both humans and mice, an upstream ERV insertion drives the primary expression of carbonic anhydrase 1 (CA1) in red blood cells (26).

Not only do ERV-LTRs supply good promoters for the host, they can also be exapted as 'ready-made' tissue-specific enhancers (18, 3). In mice, an androgen-responsive LTR upstream of the Sex-Limited Protein Gene (*Ca4/Slp*) ensures its male-specific expression (27, 31). In another case, human specific expression of salivary amylase is driven by the tissue specific enhancer effects of an upstream HERV element (32). Although both cases have proximal ERV-derived enhancers, enhancers can be distal from the genes they affect. It is likely that many ERVs have long-range enhancer effects; however, this kind of long-range effect is difficult to detect – especially with repetitive elements.

ERVs can also have a global regulatory effect through their production of and inclusion in long non-coding RNAs (lncRNAs). In a study of around 10, 000 human lncRNAs, 75-80% were found to contain TE sequences (19). Among these lncRNA-associated TE sequences, ERV-derived sequences were enriched (19). Most ERV-derived sequences in lncRNA make up 'exons', but ERVs are also overrepresented at transcriptional start sites, and particular families of LTRs have been identified as lncRNA promoters (3, 19). All in all, it seems that ERVs play a major role in the production and effect of lncRNAs. lncRNA-RoR, a HERV-H-derived lncRNA gene, has been identified as both a conserved lncRNA and as a lncRNA with broad regulatory

effects (19). In this way, ERVs – particularly ERV LTRs – can have a global regulatory effect on the transcriptome.

The effects ERVs have on a gene-by-gene scale – as promoters, alternative splice sites, transcription termination sites, transcript reducers, and enhancers – and on a global transcriptome scale – through lncRNAs and the regulation of many genes under the control of a single ERV family – have prompted some in the scientific community to refer to them as 'functional genome reshapers' (17, 33).

### 1.3.1    ERVs and Host Evolution

Through the many effects they have on a host genome, ERVs promote functional reshaping of gene regulation networks and, in so doing, promote the evolution of the host. Although they are the descendants of pathogens, the ever-growing body of evidence indicates that ERVs are necessary for the normal physiological function of their hosts and are rarely involved in initiating disease. For example, the LTR of a conserved ERV-9 insertion upstream of the β-globin gene is partly responsible for its erythroid-specific transcription, without which, primates would not have the proper proportion of haemoglobin subunits (34, 35). Simply put, this ERV, at least partly, enables normal primate respiration. Another case can be found in the human placenta (36). The *env* genes from a HERV-W and a HERV-FRD are specifically expressed in the placental tissue (37). The resultant ENV proteins, renamed syncytins, have been domesticated by the host to enable the formation of the syncytiotrophoblast (37). This layer of the placenta requires cell-cell fusion; ENV-derived syncytins enable its formation by endowing these cells with fusogenic properties (37).  Syncytins are also known to modulate immune response by donating a conserved immunosuppressive domain, thereby preventing the maternal immune system from rejecting the paternal-derived antigens of the developing child (37). In this

case, the transmembrane protein (ENV) that initially allowed these HERVs retroviral ancestors to infect the host by suppressing immune response to the virus and binding/fusing with host cells, has become critical to human reproduction (37).

Remarkably, ENV domestication for fusogenic and immunosuppressive qualities in placental tissues has happened more than once. Among eutherian mammals, mice have also had two unique ERV insertions that produce syncytin proteins (37). These proteins are quite different from the syncytins of primates, yet they perform orthologous tasks in the murine placenta (37). Some marsupials (non-eutherians) have even been found to possess unique, ERV-derived syncytins that confer the same immunosuppressive and fusogenic characteristics to their short-lived placentas (38). Indeed, the domestication of ERV-derived ENV proteins in the placenta seems to be a case of convergent evolution in mammals: multiple, independent ERV insertions have gone to fixation in different species/lineages and been exapted for the same purposes (38).

In the case of the apoptotic protein NAIP (Neuronal Apoptosis Inhibitory Protein), multiple lineage-specific ERVs have integrated upstream of the gene in humans, mice, and rats, and in each species, a set of unique ERV LTRs donate promoters to the *Naip* gene (39). This too is a case where evolution has converged on using ERV LTRs to regulate transcription of a host protein (39). At the time the mouse genome was published in 2002, Waterston *et. al* had already noted that similar TEs (even lineage-specific ones) were present in orthologous locations in human and mouse genomes (2). The fixation of similar ERVs in orthologous regions combined with the repeated exaptation of lineage-dependent ERVs to perform similar functions for their respective hosts underlines the proclivity of ERVs to confer benefits to the host. Furthermore, the clear cases of ERV exaptation-dependent convergent evolution emphasizes the importance of ERVs to genomic evolution.

ERVs have, of course, also played a major role in divergent evolution and, ultimately, speciation. Each exogenous retrovirus can only infect a certain range of hosts, and only a small fraction of exogenous retroviruses are endogenized (3). Thus, retroviral endogenization events are species and/or lineage specific. Within a lineage, species-specific expansions of a common ERV result in species-specific insertions. The differences in ERV-type and location provide variation in how the genome can be re-shaped. 34% of human-specific TE insertions have accrued within genes providing a unique and usable template for genome re-shaping (17). In fact, a comparison between humans and mice – species between which 60% of alternative promoter sequences are conserved – revealed that ERV-derived alternative promoters are almost exclusively non-conserved (18, 40). In other words, ERV-derived alternative promoters are almost always species-specific. It should be noted that the oldest ERV sequences are more likely to be shared between species and less likely to be recognized as an ERV, so this assessment is biased to find young, species-specific ERVs; however, it still provides some evidence that ERVs are involved in generating species-specific variation (18). The pattern of species-specific ERV association with differentially expressed genes is observed even among closely related primates, like chimpanzees and humans (10).

ERVs provide nearly 20% of all TF-bound transcription factor binding sites (TFBSs) in humans and mice (19, 33). TE-derived TFBSs are subject to functional constraints, yet evolve faster than non-TE-derived TFBSs – further endowing TEs with the ability to contribute to species-specific transcription units (19, 41). Among TEs, ERVs are particularly good at providing TFBSs to the human genome (41). In fact, they are the only young class of TEs that have provided the human genome with more TFBSs than would be expected based on their frequency in the genome; furthermore, those TFBSs are the most conserved of all TE-derived

TFBSs (41). This unexpectedly high conservation of TFBSs is probably due to the critical nature of some of the TFBSs that ERVs provide - OCT4 and NANOG sites, for example. At least one-fifth of these sites, which are critical for the maintenance of pluripotency and early embryonic development, reside in species-specific ERVs (27, 42). Clearly, ERVs have been and are continuing to be important for genomic plasticity.

It has also been suggested that the epigenome – the chemical modifications that dictate chromatin state, and thus provide a kind of 'oversight' to the transcriptome – was formed in response to TEs (23, 43). DNA methylation is one of the main chemical modifications of the epigenome. When transfected into cells, TEs have been shown to attract *de novo* DNA methylation (43). *In vivo*, the promoters of most viral sequences and TEs (including ERVs) are inactivated by methylation in differentiated cells (10, 24, 43). Consistent with this observation, hypomethylation of the genome – whether induced through physiological processes like aging, pathological processes like cancer, or chemical treatments like 5-azacytidine (a de-methylation agent) – de-represses ERVs, significantly increasing the number of ERV-derived transcripts present in a cell (10, 24). DNA methyltransferases and some of their associated proteins have also been shown to be necessary for ERV methylation and repression (10, 24, 44). However, in stark contrast to differentiated cells, in the pre-implantation embryo (and mouse embryonic stem cells) DNA methylation has been shown to be neither necessary nor sufficient for ERV repression (44). In these cells, histone modifications (the other major component of epigenetic control) repress ERVs (44). Both the repressive histone modifications themselves and functional copies of the proteins that are associated with depositing those modifications are necessary for ERV repression in these cells (45-47). ERVs can also be marked with repressive or active histone modifications in differentiated cells (27, 48) While all active promoters – ERV-derived

or not – are typically associated with H3K4me3, different classes of ERV are associated with different kinds of repressive modifications (*i.e.* H3K9me3 represses Class 1 and Class 2 ERVs, while H3K9me2 represses Class 3 ERVs) which suggests that these repressive modifications have been tailored to specific retroviral invasions (3, 49). Furthermore, the strong correlation between the expansions of tandem zinc finger genes and ERV invasions strongly suggests that ERVs drive the amplification and diversification of zinc finger proteins (ZFPs)(50). In mammals, most ZFPs have an associated KRAB-domain (50). KRAB-ZFPs are proteins whose job is to identify regions that require repression and recruit the necessary proteins to form repressive histone modifications at those sites (50, 51). Altogether, these lines of evidence have led to the hypothesis that the formation of the epigenome was driven, at least in part, in response to TEs and particularly, in response to ERVs (43).

Regardless of how the epigenome arose, it is certainly involved in the repression and activation of ERVs in genomes today. This interaction between ERVs and the epigenome provides an interface through which the transcriptome can respond to environmental conditions (27). ERVs and the epigenome are both sensitive to environmental stress (27, 52, 53). In fact, one third of metastable epialleles reported by Rakyan *et al.* are known to occur on ERV sequences (53). The classic case of a metastable epiallele – the A$^{vy}$ allele at the *Agouti* locus in mice – is an instance in which the interface of an ERV and the epigenome can produce an environment-sensitive phenotypic effect on the host (54). More recently, particular ERV families (like MER41B) were identified as major players in the regulation of innate immunity (55, 56). These ERVs confer the ability to produce an inflammatory response by providing regulatory elements with the TFBSs required for interferon-induced transcription proximally to important innate immunity genes throughout the genome of many mammals (55). ERVs, whether the site

of a metastable epiallele or simply acting as a *cis*-regulatory element, may have a particular proclivity to enable a genomic response to environmental conditions.

By reshaping the host genome, ERVs provide innovations that are selected for in both convergent and divergent evolution. ERVs also promote the evolution of the host genome through donating TFBSs, driving the development of the epigenome, and responding to environmental conditions – especially stress. Therefore, understanding ERVs – their regulatory capacities and their evolution within a host – is critical to understanding genome biology.

## 1.4    Intracisternal-A Particles

Intracisternal-A Particles (IAPs) are a young murine-specific superfamily of ERVs. IAPs are extremely insertionally polymorphic in mice, providing a model in which the effect of an IAP can be clearly examined due to its presence in a given location in one strain of mouse and its absence in the same location in a closely related strain of mouse (57, 58). Solitary IAP LTRs are present in highest copy number in the genome, comprising 52% of IAP sequences in the inbred strain C57BL/6 (59). Of the full-length forms, the undeleted, 7.2kb IAP sequence is present in highest copy number after solitary LTRs and the deleted, 5.4kb 1delta1 IAP sequence is present in second highest copy number (Figure 1.3) (59). The 1.9kb deletion in the internal region of the 1delta1 sequence creates a gag-pol fusion protein that aids in retrotransposition of the IAP – likely driving the 1delta1 deletion to be the most abundant of the four deletion-type IAPs (Figure 1.3) (60, 61). IAPs are particularly interesting because they experience high levels of transcription across many cell types (including many differentiated cells, not just ESCs) and still retrotranspose under normal physiological circumstances (10). Indeed, together with the ETn (early transposon) ERVs in mice, they are responsible for most new germ line ERV insertions causing phenotypes in inbred mouse strains (10). This is in stark contrast to ERV activity in

humans where ERVs are considered retrotranspositionally 'dead' (14). Although it should be noted that most of the IAP retrotransposition events in the germ line giving rise to inherited phenotypes have occurred on a single genetic background: C3H (10).

Documented, new IAP insertions typically cause ectopic gene expression from the LTR promoter (9, 10). Such a proclivity to donate promoter sequences is not common to all ERVs (10). For example, new insertions of ETns typically disrupt gene expression by donating polyA or splice sites but rarely by donating promoters, likely due to the fact that ETns are only transcriptionally active in early embryos (10). Altogether, IAPs are a particularly interesting class of ERV to study.

### 1.4.1    Long Terminal Repeats (LTRs)

LTRs have three major structural components U3, R, and U5 (from 5' to 3') (61-64). These regions are defined by the structure of the mRNA transcript (61-64). The boundary between U3 and R is defined by the transcription start site (TSS), so that the 5' end of the mRNA contains no U3 sequence but all of R and U5. The boundary between R and U5 is defined by the transcription termination site, so that the 3' end of the mRNA contains no U5 sequence but all of U3 and R (61-64). The U3 region confers all (or nearly all) promoter and enhancer properties to the LTR, and generally includes an enhancer component upstream of a CCAAT-sequence that is upstream of a TATA-related sequence (62, 65, 66). U3 regions contain a myriad of enhancer and promoter sequences, some are specific to a particular LTR, others (like SP1, GC-boxes, and YY1) are common to many LTRs – although they may vary in position within U3 (67-69).

Transcripts from full length ERVs are not prematurely terminated in the 5'LTR – even though the poly(A) sequence and the TTS are present downstream of the TSS – indicating that LTRs can be inefficient at transcription termination. Temin suggested that R-loop formation

between the U3 and R sequences was necessary for effective termination of the pro-viral transcript (64). If poly(A) termination efficiency without R-loop formation is as low as Temin believed, mRNA transcripts (which lack U3 in the 5' end) would not be able to terminate until the 3'LTR (64). Indeed, recent studies have revealed that even with R-loop formation, 3'LTRs can be inefficient at transcription termination, often resulting in very long RNA transcripts (70).

## 1.4.2 IAPLTR1 Subclasses

LTRs have been classified by sequence similarity and association with particular internal sequences (59, 71). 16.7% of the IAPLTRs in the C57BL/6 genome are IAPLTR1 (Figure 1.4) (59). This third largest class of LTRs is the youngest, as determined by the sequence divergence between LTRs of a full-length element. In fact, almost all 'perfect pair' IAPLTRs (LTRs of a full-length element with 100% sequence identity to one another) are derived from two IAP LTR classes – 1 and 1a (59). Since LTRs are identical immediately post-retrotransposition, this also implies that IAPLTR1 is one of the two classes that have had the most retrotranspositionally active IAPs recently in C57BL/6 (59). IAPLTR1s are also the most highly associated LTR with the prolific, deleted, 1delta1 IAP subtype, which is believed to have caused most of the recent retrotransposition cases in mice (59). Thus, IAPLTR1 is a particularly interesting class of LTR.

Recently, it has been suggested that IAPLTR1, along with several other subtypes of LTR including 1a, 2, 2a, 2b, 3, and 4, should actually be classified into two subclasses: a subclass that shows high divergence from the Rep-base consensus sequence (H1) and a subclass that shows low divergence from the Rep-base consensus sequence (L1) (59). Unlike the subclasses for LTRs 2, 2a, and 2b which do not correlate well with their divergence rates, IAPLTR1s H1 and L1 subclasses correlate well with their assigned divergence rates, suggesting that IAPLTR1 sequences arise from a common ancestral sequence (not independent invasions, but various
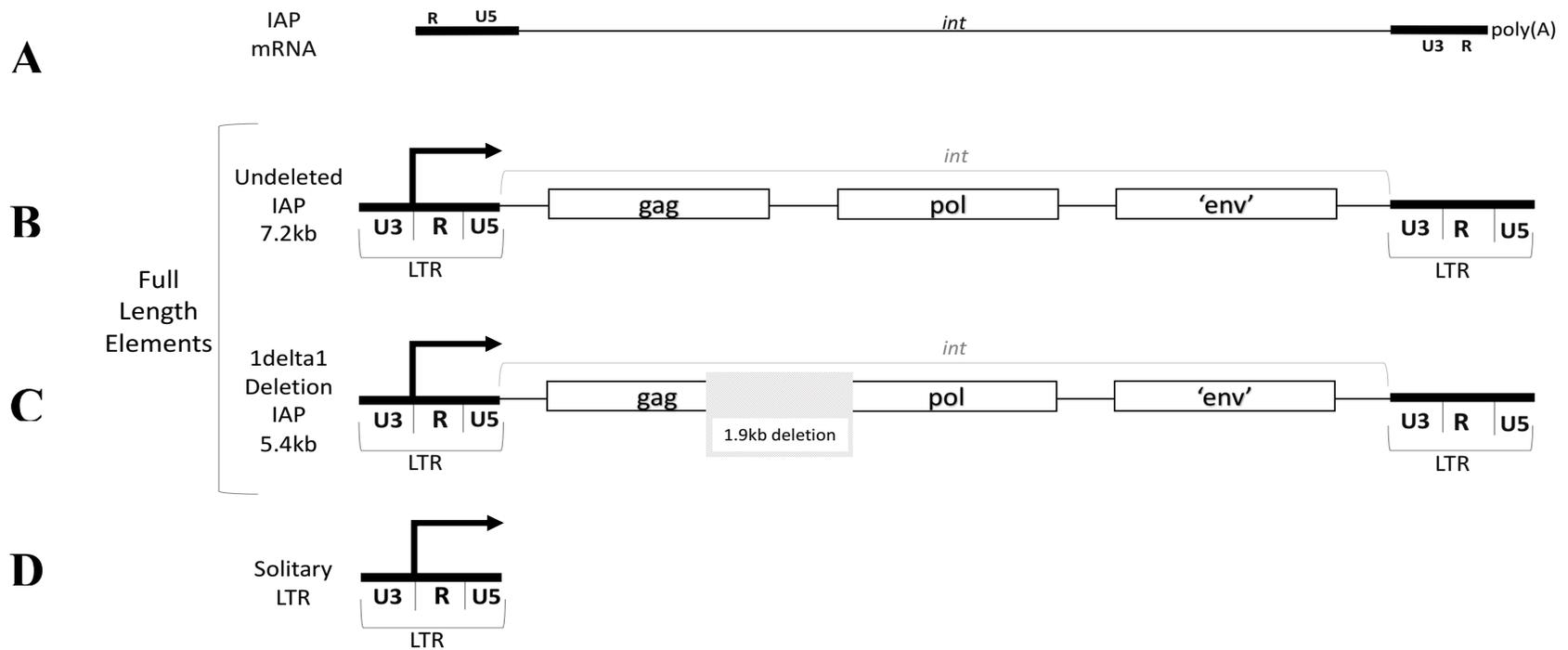
**Figure 1.3. Schematic of Main IAP Forms in the Genome**.

**(A) The IAP mRNA Transcript** defines the U3, R, and U5 regions of the LTR. **(B) The Undeleted Full-Length IAP** is 7.2kb long and contains *gag*, *pol*, and an incapacitated *env* genes. This is the form present in second highest copy number in the genome. **(C) The 1delta1 Deleted Form of the Full-Length IAP** is 5.4 kb long and contains a 1.9kb deletion in the *gag* and *pol*. This deletion creates a *gag-po*l fusion product that aids in the replication of this IAP form. It also has an incapacitated *env*. This is the form in third highest copy number in the genome. **(D) The Solitary LTR** has no internal IAP sequence. This is the form present in highest copy number in the genome. Thick black bars are LTRs, arrows are TSSs, white boxes are coding regions, 'int' specifies ERV internal region, grey is deletion

16
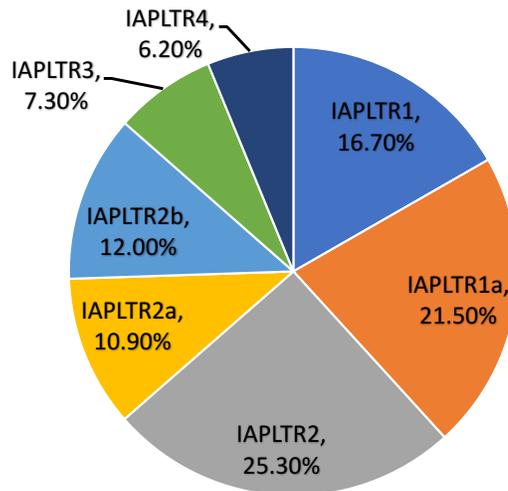
**Figure 1.4 Percent Composition of IAP LTR Classes in the C57BL6 Genome.**

[Data from reference 59]

expansions) (59). Qin *et al.* have suggested new consensus sequences for the H1 and L1 groups

(59).

### 1.4.3    IAPLTR Antisense/Bidirectional Activity

In many of the cases in which IAPLTRs have donated an alternative promoter to a host

gene, the LTR has been antisense to the gene – driving transcripts from an antisense promoter

(10). Since sense promoter activity is what drives replication and success of the ERV, it can be

assumed that any LTR with antisense-promoter effects can also drive transcription in sense,

thereby making it a candidate bidirectional promoter (as long as there has been no significant

degradation of the LTR which could eliminate promoter activity in the sense orientation). It has

long been known that some IAPLTRs have bidirectional promoter ability (65). The promoter

regions of two specific copies, IAP81 and MIA14 LTRs, were studied by using restriction sites

to clone various portions of the LTR into a reporter construct (65). This early study indicated that unique regions of these LTRs were responsible for driving sense and antisense transcription (65).

## 1.5    Bidirectional Promoters

Bidirectional promoters are generally defined as promoters for gene pairs arranged in a head-to-head orientation with less than 1000bp between transcription start sites (TSSs); the region between the TSSs is designated the bidirectional promoter (72). Further, to qualify as truly 'bidirectional', this promoter must have less than a four-fold difference between sense and antisense transcription in a given system (72). Contrary to popular belief, bidirectionality is not a property of all promoters (72, 73). In fact, bidirectional promoters bind a specific bidirectional suite of transcription factors (TFs) and, even though 78% of vertebrate TFs are underrepresented in them, bidirectional promoters experience an overrepresentation of particular transcription factor binding motifs, suggesting that this is a unique type of promoter with its own biochemistry and molecular signature (74-76). For example, TATA sequences are rare in bidirectional promoters – only 8% contain such a sequence, 18% would be expected by chance, and 28% of unidirectional promoters contain TATA sequences (72). The presence of a TATA-box may have been selected for in unidirectional or 'typical' promoters, but it appears to have been selected against in bidirectional promoters (72). Clearly then, in the case of the TATA-box, bidirectional promoters and unidirectional promoters have responded to different selective pressures.

Understanding these differences and uncovering the unique features of bidirectional promoters should be a priority because bidirectional promoters are critically important to the normal physiology of humans and of most higher eukaryotes (72-78). Although they can be up to 1000bp in length, these discrete functional units usually incorporate all of their promoter capacity into less than 300bp that are relied upon for coordinated expression of many important

genes (72). Furthermore, dysregulation of bidirectional promoters has been linked to pathologies, like somatic cancers (77). Therefore, understanding the biochemistry and molecular signature of bidirectional promoters is also critically important to furthering our understanding of physiology, pathology, and evolutionary biology.

## 1.6 Thesis Objectives

As drivers of epigenome development, re-shapers of the genome, re-wirers of gene networks, and genomic responders to environmental conditions, ERVs have a high functional impact in mammalian genomes yet they are not well understood. IAPs are a good model for ERV research because they are a young, diverse, insertionally polymorphic, species-specific, superfamily of elements that are still active in mice. Their compact promoter-enhancer regions (LTRs) have long been known to have bidirectional capacity. Even though bidirectional promoters are known to be highly important regulators with potent functional impact, these bidirectionally-endowed LTRs have not had their functionality investigated since 1988 (65). Many technologies to improve promoter studies have been published and made available since then, including: the usage of luciferase reporters instead of radiologic reporters; a noise reduction in reporter vectors; the construction of bidirectional reporter vectors; the automation of PCR and usage of high-fidelity polymerases; and many types of software and databases online that can be used for phylogenetic analysis, transcription factor binding site prediction, conservation analysis, etc... (65, 78, 124). I will apply these new technologies to:

1. Characterize the bidirectional promoter activity of IAPLTR1 and its subclasses:

    a. Test whether bidirectional reporter vectors should be used instead of unidirectional reporter vectors for assaying bidirectional promoters;

    b. Investigate whether IAPLTR1s are bidirectional promoters;

c. Identify sense and antisense core promoters present in IAPLTR1;

d. Test whether the sequence differences between H1 and L1 subclasses of IAPLTR1 that occur within the identified sense and antisense core promoters are responsible for their difference in promoter strength.

2. Investigate the evolution of IAPLTR1's bidirectional promoter activity and its H1 and L1 subclasses:

a. Evaluate whether the suggested sub-classing of IAPLTR1 based on sequence divergence has functional relevance;

b. Propose an evolutionary model for H1 and L1 subclasses;

c. Propose an evolutionary model for antisense LTR activity.

# Chapter 2: Materials and Methods

## 2.1　PCR Conditions

### 2.1.1　Phusion Polymerase

Phusion high fidelity polymerase was used according to manufacturer instructions with 5ng of template DNA and 0.5ng of template DNA and cycling as follows: 1x [98$^O$C/30"], 30x [98$^O$C/10",62$^O$C/20", 72$^O$C/15"], 1x[72$^O$C/7'], [4$^O$C/infinte].

### 2.1.2　BestTaq Polymerase

BestTaq high fidelity polymerases was used according to manufacturer instructions with 0.5ng of template DNA and cycling as follows: 1x [94$^O$C/3'], 35x [94$^O$C/10",62$^O$C/20", 72$^O$C/15"], 1x[72$^O$C/5'], [4$^O$C/infinte].

## 2.2　Restriction Digests

All restriction digest recipes were mixed well, spun down for 8sec, and incubated for 2 hours at 37C.

### 2.2.1　NcoI and XmaI Recipe

2.5μL CutSmart buffer, 0.25μL NcoI enzyme, 0.25μL XmaI enzyme, <1μg of DNA, brought to 25μL with dH2O.

### 2.2.2　NotI and NcoI Recipe

2.5μL CutSmart buffer, 0.25μL NotI enzyme, 0.25μL NcoI, 50ng DNA, brought to 25μL with dH2O.

### 2.2.3　KpnI Recipe

1μL KpnI enzyme, 2μL ReAct4 Buffer, <μg DNA, brought to 20μL with dH2O.

### 2.2.4 BglII Recipe

1200ng of DNA, 5.30μL BSA, 5.30μL ReAct3 Buffer, brought to 22μL with dH2O, 1.08μL BglII enzyme.

### 2.3 Klenow Treatment

270ng of DNA fragment, 5μL of 10mM dNTPs, 3.7μL ReAct2 Buffer, and 1.5μL of diluted Klenow were mixed and incubated at room temperature for 15min.

### 2.4 A-Tailing

5μL Reaction Buffer, 1.5μL MgCl2, 1.0μL dATP, and 5μL Taq Polymerase were added to 75ng of DNA fragment, mixed, and incubated at 70C for 30min.

### 2.5 Gel Purification

Fragments were separated in a 1.0-1.5% agarose gel, in 1xTBE with 132-157V of current for 45-120min (ranges given depend on size of fragment and agarose percentage in the gel). The gel was stained in a Gel Red solution for 20-60min and bands were visualized with UV. The appropriate-sized bands were cut from the gel with a scalpel and stored each in their own Epindorf tube. The DNA was purified from the gel with Quaigen's MiniElute Kit used according to manufacturer instructions.

### 2.6 PCR Clean-Up

QIA-Quick Spin kit was used according to manufacturer's instructions.

### 2.7 Annealing Oligonucleotides Protocol

Oligos were resuspended in annealing buffer (10mM Tris, pH 7.5-8.0, 50mM NaCl, 1mM EDTA) and mixed in equimolar concentrations. 2μg of each oligo were added to a PCR tube and brought to a total volume of 50μL in annealing buffer. The tube was heated at $95^{O}C$ for 2min in a pre-heated thermocycler and gradually allowed to cool to $25^{O}C$ over the course of 45min.

## 2.8    Plasmids

Promega's pGL3 vector was already present in the lab stocks. pLucRLuc was generously sent from David Riesman (78). Ectopic cJUN expression was driven from the Flag-JunWT-Myc plasmid which was a gift from Axel Behrens (Addgene plasmid # 47443) (120). Ectopic TFE3 expression was driven from the pEGFP-N1-TFE3 plasmid which was a gift from Shawn Ferguson (Addgene plasmid # 38120) (121). Ectopic MYC expression was driven from the pWPXL-c-Myc plasmid which was a gift from Bob Weinberg (Addgene plasmid # 36980) (122). Ectopic KLF4 expression was driven from the pWPXL-Klf4 plasmid which was a gift from Bob Weinberg (Addgene plasmid # 36981) (122). Ectopic HIF-1alpha was driven from the pcDNA3 mHIF-1α MYC plasmid (P402A/P577A/N813A) which was a gift from Celeste Simon (Addgene plasmid # 44028) (123).

## 2.9    Ligation Protocol

The following were added to a PCR tube: 2.0µL of 5x buffer, 0.5µL of ligase enzyme, linearized construct and insert were added at a 1:3 ratio as determined by the following equation:

$$required\ mass\ of\ insert\ (ng) = \frac{kb\ of\ insert}{kb\ of\ vector} \times ng\ of\ vector \times ratio\ of\ insert{:}vector$$

and brought to a total volume of 10.0µL with dH2O. The ligations were mixed by pipetting up and down five times. They were allowed to sit at room temperature for 1 hour before being moved to a $4^0$C fridge overnight.

## 2.10   Transformation Protocol

Transformations were carried out with competent DH5α cells. These cells were removed from the $-135^o$C freezer and thawed on ice. 5µL of each mini-prepped construct were added to 100µL of the competent DH5α cells, tubes were mixed gently and incubated on ice for 25min. Then the cells were heat-shocked for 45sec in a water bath at exactly $24^o$C before the tubes were

returned to ice for 2min. After 2min on ice, the cells were rescued with 800µL of room temperature LB (antibiotic-free) and moved to a 37°C shaker at 110rpm for 1hr30min.

The entire bacterial culture was poured into a micro-centrifuge tube and centrifuged at 5000rcf for 1min to pellet the cells, ~600µL of supernatant was poured off. The cells were gently pipetted up and down to re-suspend them in the remaining 200µL of liquid. This solution was plated on an LBAmp$^+$ or LBKan$^+$ (depending on what is appropriate) plate under sterile conditions and spread with a hockey stick. These plates were allowed to dry right-side up, then were turned upside down and moved to the 37°C incubator overnight. Once colonies have grown up, the plate was sealed with parafilm and moved to the 4°C fridge.

## 2.11  Overnight Cultures

A single bacterial colony was picked with an autoclaved toothpick, touched to an appropriately labelled region of a new LBAmp$^+$ or LBKan$^+$ plate, and dropped into a pre-dispensed 1.4mL LBAmp$^+$ (or LBKan$^+$). The tooth-pick containing-LBAmp$^+$ tube was then incubated at 37°C overnight with shaking at 150rpm. The plate was turned upside down and incubated at 37°C overnight.

## 2.12  Glycerol Stocks

40% glycerol (w/v) in dH2O was prepared and autoclaved. 0.5mL of an overnight bacterial culture was pipetted into a cryo-vial. 0.5mL of the prepared glycerol was added to the tube. The mixture was then pipetted up and down gently to mix, and the bacterial glycerol stock was moved to -70C for storage.

## 2.13  Mini-Prep Protocol

The 1.4mL of liquid bacterial culture from an overnight was poured into a micro-centrifuge tube and pelleted at 5000rcf for 1min. The supernatant was poured off and the cells

were re-suspended in 600μL of dH$_2$O. The Zippy$^{TM}$ Plasmid Miniprep Kit by Zymo Research

was then used according to the manufacturer's instructions to separate the plasmid from the *E.*

*coli*. The plasmid was eluted through the column with 20μL of dH$_2$O.

**2.14  Maxi-Prep Protocol**

A 150mL bacterial culture was grown up overnight at 37$^o$C with shaking at 150rpm in

LBAmp$^+$ solution. The solution was then poured into a large, plastic tube. The tube was weighed

and another tube was filled with dH$_2$O until it weighed within 1g of the tube containing the

bacterial culture. The tubes were placed in the JA10 rotor for the centrifuge so that it would be

balanced, and the cultures were pelleted by spinning at 6000rcf for 15min at 4$^o$C. Following

centrifugation, the supernatant was poured off and the tube was left open and inverted on a paper

towel to dry the pellet. 12mL of S1 Buffer were added to the pellet, which was then vortexed and

shaken to break up the cells into a uniform suspension. 12mL of S2 Buffer were then added to

the solution. The tube was capped, and inverted 6-8 times for mixing. 12mL of S3 Buffer was

added less than five minutes after the addition of the S2 Buffer, the tube was re-capped and

mixed by inverting another 6-8 times. The capped tube was incubated on ice for 5min. While the

incubation was happening, the filter paper was fluted and placed 1:1 into plastic funnels. Two

large test-tube racks were stacked and a large tube was placed in the lower rack with a column

that has had 6mL of Buffer N2 flow through it prior. The funnels were then placed in the

columns. After the solution had incubated for 5min on ice, it was poured out into the fluted filter

paper in a plastic funnel. The flow through was poured out and the large tube was replaced with

a new 50mL conical. 32mL of Buffer N3 was washed through into the 50mL conical for about

15min, where it was measured. While Buffer N3 was cleaning the DNA on the column, the large

centrifuge tube was washed with ethanol, poured out, and let sit upside down on paper towel

until dry (about 30min). The centrifuge tube, once clean and dry, was placed under the column

and the DNA was eluted off the column with 15mL of Buffer N5. 11mL of 100% isopropynol

were added, the tube was capped and mixed by inverting 3-4times. The tube was weighed, and

100% isopropynol was added until all tubes were within 1g of each other. The tubes were placed,

un-capped in the JA20 rotor for the centrifuge so that it would be balanced, and the cultures were

pelleted by spinning at 15, 000rcf for 30min at $4^{o}$C. The isopropynol was then poured off. 5mL

of 70% ethanol were added to the DNA pellet, the tube was swirled to wash the pellet with the

ethanol. The tubes were placed in the JA20 rotor for the centrifuge so that it would be balanced,

and the cultures were pelleted by spinning at 15,000rcf for 10min at room temperature. The

ethanol was pulled off with a pipette, and the pellet was left to dry overnight. 300µL of $dH_2O$

were added to the pellet and the tube was swirled until the pellet was fully dissolved (until the

water no longer appears to 'stick' to the pellet). This solution was then transferred to micro-

centrifuge tubes.

## 2.15 IL3 LTR Fragment Constructs (pLucRLuc)

IL3 LTR fragments were generated with XmaI and NcoI overhangs (the primers for the

fragment generation are listed in the table), and pLucRLuc was linearized with XmaI and NcoI.

Each LTR fragment was gel purified and ligated into gel purified linear pLucRLuc.

## 2.16 Point Mutations Constructs (pLucRLuc)

QuikChange Lightening Site-Directed Mutagenesis Kit by Agilent Technologies was

used according to manufacturer instructions in order to induce point mutations in IAPLTR1

subclass representatives. Primers for point mutations were designed using Agilent Technology's

QuikChange Primer Design Program. The reaction mix was mixed as follows (this recipe is for a

single reaction): 5µL 10xreaction buffer, 1.66µL of 30ng/µL template plasmid, 1.25µL of

100ng/μL for each Primer, 1.0μL dNTP XLMix, 37.3μL of dH$_2$O, 1.5μL QuikSolution, and 1.0μL of enzyme. After the reactions have been mixed, cycling was performed to the following specifications: 1x [95$^o$C/2"], 18x [95$^o$C/20',60$^o$C/10', 68$^o$C/3"], 1x[68$^o$C/5"], [4Cinfinte].

2μL of DpnI enzyme were added directly to each amplification reaction after the cycling had been completed. The mixture was pipetted gently up and down 20 times, flicked, spun down for 6sec, and incubated in a 37$^o$C water bath for 5min. Following the DnpI digestion, these solutions were transformed into competent DH5α cells according to protocol 2.9.

## 2.17   Cell Culture

### 2.17.1   Freeze-Medium Recipe

1mL of fresh freeze medium was prepared for each cryotube. FBS was mixed with 1:10 DMSO, then filtered through a 0.22G (or smaller) sterile filter and syringe. Freeze medium was used cold.

### 2.17.2   Thawing Cells

9mL of warm media were added to a 15mL conical tube. Cells were rapid-thawed from the -135$^o$C freezer by holding the cryotube in a 37$^o$C water bath and swirling gently until only a small lump of ice was free-floating inside. The tube was then moved to the BSC where a P1000 was used to remove all the cells (the full 1mL of liquid). The cells were gently expelled into the 9mL of pre-allocated warm media before being spun down to a pellet (300G for 5min). The supernatant was pulled off the pellet and discarded. The pellet was re-suspended in 1mL of media and mixed well before another 4mL of media were mixed in. Cells were counted and the appropriate number of cells was plated.

### 2.17.3   p19 Growth Medium Recipe

DMEM with 10% FBS and 100x (1:100) Penecillin/Streptomycin.

### 2.17.4   p19 Cell Culture Passaging

The cells were trypsinized with 2mL of Trypsin-0.25%EDTA and tapping until the cells were all drifting free from the surface of the 100mm$^2$ culture plate (visualized under the microscope). The Trypsin-0.25% EDTA was then deactivated by the addition of 8mL of p19 growth medium. All 10mL of liquid were pipetted into a 50mL conical tube and pelleted by centrifugation for 5min at 300rcf. The supernatant was then pulled off and discarded and the pellet was re-suspended in 1mL of fresh p19 growth medium. The liquid was pipetted gently up and down with the p1000 40 times, then again with the p200 to break up the cells. 1mL of the cell-suspension was diluted into 9mL of fresh p19 growth medium. The diluted cell suspension was then pipetted into a 100mm$^2$ culture plate, that was subsequently incubated at 37$^o$C and 5% $CO_2$. This protocol was followed every 2-3 days.

### 2.17.5   p19 Cell Seeding of a 24-Well Plate

The cells were trypsinized with 2mL of Trypsin-0.25%EDTA and tapping until the cells were all drifting free from the surface of the 100mm$^2$ culture plate (visualized under the microscope). The Trypsin-0.25% EDTA was then deactivated by the addition of 8mL of p19 growth medium. All 10mL of liquid were pipetted into a 50mL conical tube and pelleted by centrifugation for 5min at 300rcf. The supernatant was then pulled off and discarded and the pellet was re-suspended in 1mL of fresh p19 growth medium. The liquid was pipetted gently up and down with the p1000 40 times, then again with the p200 to break up the cells.

10μL of the cell-solution was mixed in a microcentrifuge tube with 10μL of Trypan-Blue. The tube was removed from the cell hood and 10μL of the Trypan-Blue solution was injected into either side of a hematocytometer. The live-cells were then counted and the following equation was used to calculate the number of live-cells per mL:

$$\frac{number\ of\ live\ cells}{number\ of\ squares\ counted} \times 20\,000 = \frac{number\ of\ live\ cells\ in\ solution}{mL\ of\ solution}$$

The number of cells required for seeding a plate at 60, 000 cells per well was found with the following equation:

$$60\,000\ cells \times number\ of\ wells = number\ of\ cells\ required$$

The amount of cell solution required was found using this equation:

$$number\ of\ cells\ required \times \frac{mL\ of\ solution}{number\ of\ live\ cells\ in\ solution}$$

The amount of cell-solution necessary was aliquoted and brought to volume with p19 growth medium. The volume required was 1mL per cell well that needs seeding. Once the diluted cell solution was well-mixed, the 10mL pipet was used to add 1mL of solution dropwise to each well. The plate was covered and moved front-to-back and side-to-side to ensure even distribution of the cells in each well. Then the plate was incubated at 37$^{\circ}$C and 5% $CO_2$ until the wells were 60-80% confluent (about 24hrs of growth).

## 2.18  Transfections

600ng of each construct and 300ng of pCMV-βgal and, in the case of the overexpression assays, 300ng of the ectopic expression vector were brought to volume of 75µL with Opti-MEM and left to incubate at room temperature while the next step was completed. Lipofectamine 2000 was mixed with Opti-MEM at a ratio of 1:25 (Lipofectamine2000:Opti-MEM); this solution was left to incubate at room temperature for 5min. 50µL from the DNA-Opti-MEM solution was mixed with 50µL from Lipofectamine2000-Opti-MEM solution, and left to incubate at room temperature for 20min. While the solutions were incubating, the seeded 24-well plate was removed from the 37$^{\circ}$C incubator, the growth medium was removed from each well and discarded, the cells were washed with 500µL of serum-free, antibiotic-free DMEM, and 500µL

of serum-free, antibiotic-free DMEM were added to each well. 50μL of each incubating solution were added dropwise to each well. Technical replicates were performed for each solution.

## 2.19  Lysate Preparation

Before lysing the cells, all growth medium was pulled off and discarded. The cells were washed twice with 500μL PBS, and 100μL of 5x Passive Lysis Buffer (from the Dual Luciferase Kit by Promega). The cells were lysed in this buffer, with shaking, for 15-25min at room temperature. The wells were scraped, underwent freeze-thaw (at least 30min at -20C, with thawing at room temperature), and scraped again. The lysates were pipetted vigorously up and down and transferred to epindorf tubes. Lysates were stored in Styrofoam boxes at -20C.

## 2.20  Luciferase Measurements

50.0μL of LarII were pre-dispensed into round-bottom polystyrene tubes. 5.0μL of well-mixed lysate were added and the reaction was measured immediately in a luminometer. 50.0μL of Stop'N'Glo were added, pipetted gently up and down 5-10 times, and measured in a luminometer.

## 2.21  Galactosidase Measurements

300.00μL of galactostar were pre-dispensed into round-bottom polystyrene tubes. 5.0μL of lysate were added and mixed. The lysates were incubated in galactostar for 90-110min before being measured in a luminometer.

## 2.22  Statistics

### 2.22.1  Percentage of Full Length Activity for Luciferase Assays

Raw luciferase measurements, whether Firefly or Renilla, were normalized by their respective raw galactosidase measurements using the following equation:

$$\frac{Raw\ Luciferase\ Measurement}{Raw\ Galactosidase\ Measurement} = Normal\ Luciferase$$

The mean of normalized technical replicates was found:

$$\frac{(Normal\ Luciferase\ 1 + Normal\ Luciferase\ 2)}{2} = Biological\ Replicate$$

This is the biological replicate for luciferase activity. An experimental biological replicate (*i.e.* any value ascertained with pLucRLuc containing an altered version of a full length LTR), was then divided by its according control biological replicate (*i.e.* full length, unaltered IL3LTR or Wnt9bLTR), and converted to a percentage of full-length promoter activity:

$$\frac{Experimental\ Biological\ Replicate}{Control\ Biological\ Replicate} \times 100\% = \%\ Full\ Length\ Activity\ Biological\ Replicate\ 1$$

The mean of the percentage of full-length promoter activity for each experimental value was found:

$$\frac{\Sigma(\%\ Full\ Length\ Activity\ Biological\ Replicates)}{Number\ of\ Biological\ Replicates} = Mean\ \%\ Full\ Length\ Activity$$

The mean percentage of full length activity was the final value found and the value that was represented in Figure graphics.

**2.22.2   Error Bars for Percentage of Full Length Activity Luciferase Assays**

The standard deviation of Biological Replicates was found using the following formula (TR = Technical Replicate):

$$Standard\ Deviation = \sqrt{\frac{(TR\ 1 - Mean\ of\ TRs)^2 + (TR\ 2 - Mean\ of\ TRs)^2}{2 - 1}}$$

This number was then represented as a Percentage of Full Length Activity for each biological replicate (BR):

$$Percentage\ Error\ for\ a\ given\ BR = \frac{Standard\ Deviation}{Mean\ of\ Control\ TRs}$$

When the Mean % Full Length Activity was found, the Percentage Error associated with each biological replicate was propagated using the following formula to find the Final Error:

$$Final\ Error = \sqrt{(Percentage\ Error\ for\ BR\ 1)^2 + (Percentage\ Error\ for\ BR\ 1)^2}$$

Final Error was represented by error bars in the figure graphics. If the final number of units is not represented as a percentage (*i.e.* If it is relative luciferase units), then final error is found by skipping the calculation for percentage error.

### 2.22.3 Calculations for Statistical Significance

The two-sample, two-tailed T-test was used for all calculations of statistical significance. This was the appropriate test because all comparisons were made between the means of two independent populations and the means could be greater or lesser than each other (Biological statistics text). In order to calculate t-value, the following equation was used:

$$t = \frac{Mean\ \%\ Full\ Length\ Activity\ 1 - Mean\ \%\ Full\ Length\ Activity\ 2}{\sqrt{\frac{(Standard\ Deviation\ 1)^2}{2} + \frac{(Standard\ Deviation\ 2)^2}{2}}}$$

The degrees of freedom (DF) present in this system were found by using the equation:

$$DF = number\ of\ measurements\ per\ Mean\ \%\ Full\ Length\ Activity - 1$$

Table A.2 of Introductory Biological Statistics (126) was then used assign significance to comparisons of sample means by finding the critical value of t at one degree of freedom for any significance greater than or equal to p=0.05.

### 2.23 Transcription Factor Candidate Selection

The three regions of interest from IL3LTR and Wnt9bLTR (in appendix), as well as the full LTRs, were analyzed for TFBS by PROMO-Alggen and oPPOSUM3.0.

### 2.23.1   PROMO-Alggen

This software uses a two-step process for submission of sequences for TFBS analysis. In step one, 'SelectSpecies' and 'SelectFactors' were set to *Mus musculus*. In step two, the sequences for analysis were passed in FASTA format into the 'STRING TO BE ANALYZED' box, and the search was carried out with a maximum matrix dissimilarity rate of 15.

### 2.23.2   oPPOSUM 3.0

Sequence-based Single Site Analysis (SSA) was selected as the mode of inquiry. This software uses a four-step process for submission of sequences for TFBS analysis. In step one, the sequences for analysis were pasted in FASTA format into the submission box. In step two, a background of sequences randomly selected from experimental control peaks derived from mouse fibroblasts was selected. In step three, the JASPAR CORE profile for vertebrates was selected as the transcription factor binding matrix. Finally, in step four, the matrix match threshold was set at 85%, all results were set to be displayed, and the results were sorted by Z-score.

### 2.24  Alignments and Phylogenies

DNA alignments were performed on the SeaView software suite. The MUSCLE aligner was used allowing for gaps. Phylogenies were constructed with PhyML software, typically bootstrapped to 1000.

# Chapter 3: Results and Brief Discussion

## 3.1  Bidirectional Reporter Vectors are Better than Unidirectional Reporter Vectors for Assaying Bidirectional Promoters.

In the past, bidirectional promoter activity has been assayed by cloning a promoter of interest into a unidirectional reporter vector in both orientations (5' to 3' and 3' to 5') (Figure 3.1). In 2011, Polson *et al.* built a bidirectional reporter vector – pLucRLuc – so that bidirectional promoter activity could be assayed by one vector in a single experiment (Figure 3.1) (78). To determine which kind of reporter vector (unidirectional or bidirectional) should be used when asking questions of bidirectional promoter activity comparisons were made between their sources of technical error, theoretical biases, and experimental performance.

## 3.1.1  Bidirectional Reporter Vectors Reduce Sources of Technical Error for Transient Transfection Experiments.

Technical error is experimental error that is attributed to variation in technique or measurement, or any variation inherent in the experiment itself (79). There are many sources of technical error in experiments involving transient transfections of reporter constructs. Table 3.1 summarizes some of these sources of error and whether they are present when using a bidirectional reporter or a unidirectional reporter.

Because unidirectional reporter constructs require the direct comparison of two transfection series – one that reports on sense promoter activity with the promoter cloned 5' to 3' with respect to the reporter and a secondary transfection series that reports on antisense promoter activity with the promoter cloned 3' to 5' with respect to the reporter (Figure 3.1) – there are many instances where technical error can be introduced between the sense and antisense measurements (Table 3.1). For example, slight differences in the cleanliness or concentration of
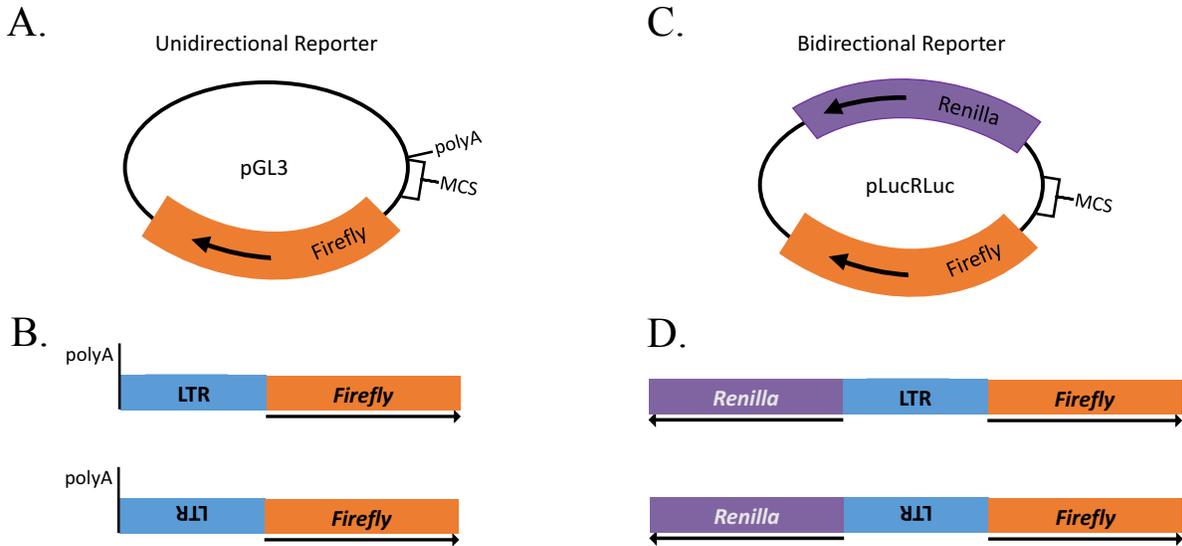
**Figure 3.1 Unidirectional and Bidirectional Reporter Vector Schematics.**

**A. Unidirectional reporter vector schematic, circular**. pGL3 is the model for this schematic [Information from reference 125]. Included in the illustration is the polyA site on one side of the multiple cloning site (MCS) and the reporter, Firefly luciferase (the black arrow indicates the orientation of the reporter), on the other. **B. Unidirectional reporter vector schematic, linear**. Illustrated here is the experimental set up with the LTR cloned into the MCS in both orientations. The upper schematic shows the LTR 5'-to-3' with Firefly luciferase to report sense transcription and a ployA site to block antisense transcription. The lower schematic shows the LTR 3'-to-5' with Firefly luciferase to report antisense transcription and a polyA site to block sense transcription. **C. Bidirectional reporter vector schematic, circular**. pLucRLuc is the model for this schematic. Included in the illustration is the MCS between the Firefly and Renilla luciferases (again the black arrow indicates the orientation of the reporter). **D. Bidirectional reporter vector schematic, linear**. Illustrated here is the experimental set up with thte LTR cloned into the MCS in both orientations. The upper schamtic shows the LTR 5'-to-3' with Firefly luciferase to report sense transcription and Renilla luciferase to report antisense transcription. The lower schematic shows the LTR 3'-to-5' with Firefly luciferase to report antisense transcription and Renilla luciferase to report sense transcription. [Information from references: 78, 125]

the DNA prep as well as variation from mixing, measurement, or administration of DNA or reagents are sources of technical error inherent in needing to directly compare two transfections. Even though transfection efficiency is controlled (so the absolute number of cells and how well they were transfected is taken into account) well-to-well variation in cell growth and maturity can still introduce error due to differences in cell type and cell function. Well-to-well variation in cell lysis can also introduce error prior to the measurement of sense and antisense promoter activity. On the other hand, when using a unidirectional reporter construct, Firefly luciferase is typically the experimental reporter and Renilla luciferase is typically the control for transfection efficiency. Renilla luciferase produces a stable bioluminescent signal that allows for consistent measurements of transfection efficiency (see the technical manual for Galacto-Star page 3, and Dual Luciferase System page 13) (80, 84). This reduces the amount of technical error expected from this measurement relative to other types of transfection efficiency reporters.

Alternatively, bidirectional reporter constructs use Firefly and Renilla luciferases as experimental measures. Renilla luciferase, therefore, cannot be used as a measure of transfection efficiency, forcing bidirectional reporter assays to use a reporter with decreased consistency of measurement and thus higher inherent error than Renilla for transfection efficiency. Typically, galactosidase is used. However, the reporter of transfection efficiency is the only point on which assays with bidirectional reporter constructs appear to be a source of more technical error than assays using unidirectional reporter constructs. For all other identified parameters, the ability to test sense and antisense promoter activity from a single DNA prep and a single well reduces sources of technical error. Therefore, using a bidirectional, instead of unidirectional, reporter construct to assay the bidirectional capacity of a promoter reduces the number of sources of technical error for the experiment.

**Table 3.1 Sources of Technical Error for Transient Transfection Experiments Using Unidirectional or Bidirectional Reporter Constructs.**

| Source of Technical Error | Unidirectional Reporter Construct | Bidirectional Reporter Construct |
|---|---|---|
| **Cleanliness of DNA Prep** | Present, difference between cleanliness of either DNA prep for either orientation | Absent, single DNA prep so no error can be introduced between sense and antisense |
| **Concentration of DNA Prep** | Present, slight differences in concentration in DNA prep can change absolute amount of DNA present for either | Absent, single prep so absolute amount of DNA tested for sense and antisense is the same |
| **Mixing Variability** | Present, differences can be introduced from slight variation in mixing of reagents or DNA preps | Absent, single prep so no differences introduced between sense and antisense as a result of mixing variation |
| **Measurement Variability** | Present, differences due to pipetting error, variation inherent to measuring the concentration of DNA in solution, etc… | Reduced, single prep so fewer differences introduced between sense and antisense as a result of measurement variability; although pipetting error during luciferase analysis could still introduce error between sense and antisense |
| **Reagent Administration** | Present, well-to-well differences in dispersion of reagents and DNA | Absent, single well so no difference in dispersion of reagents and DNA |
| **Cell Growth/Maturity** | Present, well-to-well differences in cell-cell contact and differentiation | Absent, single well so no differences in cell-cell contact and differentiation |
| **Cell Lysis** | Present, well-to-well variation in cell lysing process can vary in concentration of reporter or the amount of cells still adhering to the plate | Absent, single well so no variation between sense and antisense measurements induced from differences in lysis process |
| **Variation Inherent in Measure of Transfection Efficiency** | Reduced, Renilla luciferase (typically used) produces a highly stable bioluminescent molecule that results in highly consistent measurements | Increased, Renilla luciferase is an experimental measure and thus cannot be used, galactosidase (typically used) produces an unstable molecule that results in less consistent measurements |

**3.1.2    Bidirectional Reporter Constructs are Theoretically Less Likely to Bias the**

**Experimental Results of Bidirectional Promoter Activity than a Unidirectional Reporter**

**Construct.**

The purpose of a promoter assay is to test the capacity of given DNA to promote transcription outside of the constraints of its native genomic environment. When a bidirectional promoter is assayed using a unidirectional reporter construct, there is a 'genomic constraint' in that the vector dictates that transcription will only occur in one direction through the addition of the poly(A) site upstream of the multiple cloning site. In this way the construct itself, if unidirectional, could introduce a bias to a bidirectional promoter assay. Two examples of how this constraint could introduce a significant bias to bidirectional promoter data obtained with a unidirectional reporter construct are: 1. Polymerase stalling resulting in decreased loading efficiency, and 2. Incapacitating inherent enhancer effects of a promoter.

**3.1.2.1    Polymerase Stalling Resulting in Decreased Loading Efficiency of More**

**Polymerases.**

If a bidirectional promoter has been cloned into a unidirectional reporter construct, then polymerases recruited for promoter activity in the orientation away from the reporter could remain stalled on the promoter (81). Because these polymerases have nowhere to go, they could sit stagnant on the promoter and sterically hinder other polymerases – even polymerases for promoter activity into the reporter – from loading. Thus, unidirectional reporter constructs could induce a net decrease in polymerase loading efficiency, possibly resulting in net reduction of promoter activity.

### 3.1.2.2    Incapacitating inherent enhancer effects of the promoter.

Polymerase activity in sense and antisense directions across a bidirectional promoter and out into the proximal DNA is thought to increase accessibility of DNA and maintain open chromatin (82). Recently, enhancer RNAs (eRNAs) have been cited as important for the functionality of active enhancers although there has been some evidence to suggest that bidirectional promoter activity is different from eRNA-related enhancer activity (83). Regardless of whether a bidirectional promoter has eRNAs-related enhancer activity or simply promotes bidirectional transcription, if it is being tested in a unidirectional reporter construct, it may not promote transcripts as strongly as it would in a bidirectional reporter construct in which the polymerase was free to enter the proximal DNA in either direction (not just one). In this way, a unidirectional reporter construct could incapacitate the inherent enhancer effects of a promoter, again decreasing the net promoter activity.

### 3.1.3    Bidirectional Reporter Constructs Return Different Experimental Results than Unidirectional Reporter Constructs.

Speculation that there could be differences between unidirectional and bidirectional reporter constructs are only important considerations if there is a significant difference between data reported with unidirectional reporter and data reported with a bidirectional one. To investigate if there was a difference in measured promoter activity based on the type of reporter construct, a bidirectional promoter was assayed in a unidirectional reporter construct (pGL3) and in a bidirectional reporter construct (pLucRLuc) (78, 80, 125). However, since pLucRLuc uses Renilla luciferase as a measure of antisense activity, Renilla first had to be compared to Firefly to see if measurements taken with these luciferases are directly or indirectly (via a constant ratio of renilla:firefly) equivalent (78).

As will be further discussed in the next section, Wnt9bLTR and IL3LTR are bidirectional promoters. These promoters were cloned into both orientations (5' to 3' and 3' to 5') in pLucRLuc. Sense activity measured with Firefly was directly compared to sense activity measured with Renilla, and antisense activity measured with Firefly was directly compared to antisense activity measured with Renilla (Figure 3.2, Table 3.2). Although there is a one-to-one directly comparable relationship between antisense measurements with Renilla and Firefly luciferase for the Wnt9bLTR, this relationship is inconsistent – even for sense activity of the same promoter (Figure 3.2, Table 3.2). Where antisense activity measurements seemed directly comparable with Renilla and Firefly, sense promoter activity for Wnt9b LTR was found to have a 4.56-fold difference between Renilla and Firefly measurements (Figure 3.2, Table 3.2). Clearly then, Renilla and Firefly measurements are inconsistent in their comparability with respect to a given promoter – neither is it possible to directly compare Firefly and Renilla measurements to each other nor is it possible to establish an indirect relationship between the two. Renilla and Firefly luciferases are not equivalent in this respect.

The second test of equivalence was to see if these relationships were consistent between promoters. IL3LTR promoter activity was measured after it had been cloned in both orientations; sense Firefly activity was compared to sense Renilla activity, and antisense Firefly activity was compared to antisense Renilla activity. Although the relationship between Renilla and Firefly readings for this promoter seem to be consistently 8:1 Renilla:Firefly, IL3LTR Renilla and Firefly measurements were non-equivalent to Wnt9bLTRs Renilla and Firefly measurements (Figure 3.2, Table 3.2). Renilla had an 8-fold increase over Firefly in both sense and antisense for IL3LTR, but had a one-to-one Firefly to Renilla for Wnt9b's sense promoter activity and a 4.56-fold Renilla to Firefly for Wnt9bLTRs antisense activity (Figure 3.2, Table 3.2). Thus,

Renilla and Firefly luciferases are also inconsistent in their relationship to each other when assaying different promoters.

Conclusively then, Renilla luciferase cannot be compared to Firefly luciferase directly (with the expectation of a consistent 1:1 relationship between the luciferases) or indirectly (with a ratio that could be consistently established either between different promoters or within a given promoter).

To investigate whether there was a difference in measured promoter activity based on the type of reporter construct (unidirectional or bidirectional), measurements with Firefly luciferase could only be compared to other measurements with Firefly luciferase. So IL3LTR was cloned in both orientations into the well-characterized, unidirectional reporter construct pGL3 and the new, bidirectional reporter construct pLucRLuc. Sense and antisense promoter activity could thus be measured using only Firefly luciferase (since that is the reporter present in the pGL3 plasmid), and direct comparisons could be made (Figure 3.3). For both sense and antisense of the same promoter (the IL3LTR), the unidirectional reporter construct (pGL3) returned significantly different measurements than the bidirectional reporter construct (pLucRLuc) (Figure 3.3).

Because there is a significant difference between the measurements taken with a unidirectional reporter and a bidirectional reporter, and there are theoretical biases that could be induced by a unidirectional reporter on the measurements of activity of a bidirectional promoter, a bidirectional reporter construct should be used to assay bidirectional promoters. Moreover, the use of a bidirectional reporter construct decreases the overall sources of technical error that are inherently present in assaying bidirectional promoter activity. The pLucRLuc bidirectional reporter construct will therefore be used instead of a unidirectional reporter construct in order to assay the bidirectional promoter activity of LTRs in the following experiments.

### 3.2   IAPLTR1 Subclass Analysis

Qin *et al.* proposed that, due to sequence divergence, two subclasses of the IAPLTR1 class of mouse ERVs would be a more representative system of nomenclature than the current, single classification (IAPLTR1) (59). The two classes identified – high divergence (H1) and low divergence (L1) from the IAPLTR1 Rep-base consensus sequence – were each given a new consensus sequence that was derived only from the sequences that clustered as L1 or H1 (59). To test whether the differences in the sequence of H1 and L1 IAPLTR1 subclasses have functional relevance, recent IAPLTR1 retrotransposition cases were classed as either H1 or L1 and a representative LTR sequence was selected for each subclass. These representative cases should be functionally relevant given their recent retrotransposition and their previously characterized promoter ability. Here, retrotransposition events and promoter activities of transcriptionally active IAPLTR1s that clustered most closely with either the H1 or L1 subclasses consensus sequence were compared.

### 3.2.1   Recent IAPLTR1 Retrotransposition Cases

Recent cases of IAPLTR1 retrotransposition were identified in the literature. In order to identify representative sequences for the suggested H1 and L1 subclasses of IAPLTR1, only cases with fully sequenced LTRs could be used for this analysis. These recently retrotransposed and fully sequenced IAPLTR1s were further filtered by known promoter activity. Since the goal was to elucidate any functional differences in promoter activity associated with sequence divergence, only the IAPLTR1s that are known to make up the 5' end of a transcript were considered in this analysis. Six cases of germline IAPLTR1 insertions were identified according to the above constraints. Three of these, including $A^{iapy}$, $A^{aiy}$, and $A^{vy}$, are insertions into the *Agouti* gene (85-87). The fourth IAPLTR1 insertion is downstream of the *Wnt9b* gene (88).

A.
Renilla | LTR | Firefly

B.
IL3 LTR
1,586.60 | 70.20

C.
Wnt9b LTR
50.63 | 21.68

D.
IL3 LTR
572.17 | 201.27

E.
Wnt9b LTR
98.86 | 45.50
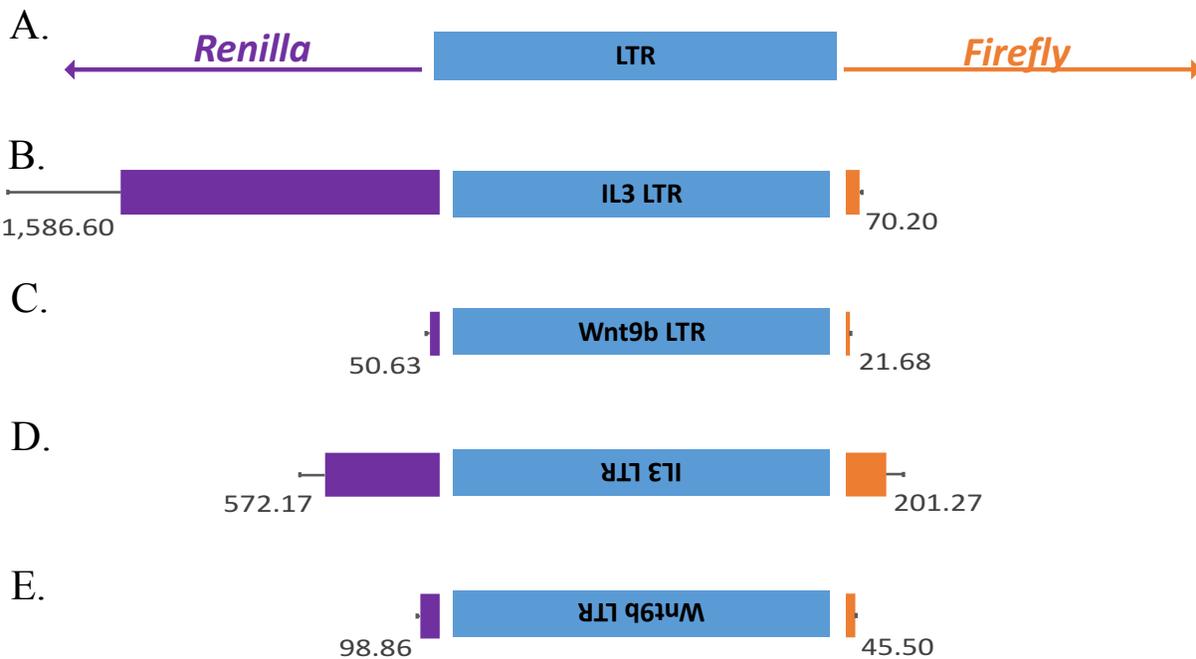
**Figure 3.2 Comparisons of IL3LTR and Wnt9bLTR, and Firefly Luciferase and Renilla Luciferase.**

**A. Schematic of and LTR cloned in between Renilla and Firefly luciferase as it is in the bidirectional promoter pLucRLuc. B-C. IL3LTR and Wnt9bLTR cloned in the sense orientation** in the bidirectional promoter pLucRLuc. **D-E. IL3LTR and Wnt9bLTR cloned in the antisense** orientation in the bidirectional promoter pLucRLuc. Renilla and Firefly Luciferases were found to be non-equivalent. Sense and antisense promoter activities were found to be within four-fold of each other for both of the IAPLTRs tested. The IL3LTR was found to have stronger sense and antisense promoter activities than Wnt9bLTR. These comparisons can be found in **Error! Reference source not found.** Purple is Renilla, Orange is Firefly, blue is LTR.

**Table 3.2 Table of Comparative Values to Accompany Figure 3.2**

| Type of Comparison | LTR | Promoter Activity Compared (S is sense; AS is antisense) | Comparative Value | Significance Value |
|---|---|---|---|---|
| Firefly (FF) vs. Renilla (Rn) | IL3 | S B. (FF) vs S D. (Rn) | 8.15-fold difference | P ≤ 0.001 |
| | | AS B. (Rn) vs AS D. (FF) | 7.88-fold difference | P ≤ 0.05 |
| | Wnt9b | S C. (FF) vs S E. (Rn) | 4.56-fold difference | P ≤ 0.05 |
| | | AS C. (Rn) vs AS E. (FF) | 1.11-fold difference | n.s. |
| Sense vs Antisense Bidirectional Promoter Activity | IL3 | S B. (FF) vs AS D. (FF) | 2.86-fold difference | -- |
| | | AS B. (Rn) vs S D. (Rn) | 2.77-fold difference | -- |
| | Wnt9b | S C. (FF) vs AS E. (FF) | 1.95-fold difference | -- |
| | | AS C. (Rn) vs S E. (Rn) | 2.10-fold difference | -- |
| Sense vs Antisense Promoter Comparison | IL3 vs Wnt9b | S B. (FF) vs S C. (FF) | -- | P ≤ 0.001 |
| | | AS B. (Rn) vs AS C. (Rn) | -- | P ≤ 0.01 |
| | IL3 vs Wnt9b | S D. (Rn) vs S E. (Rn) | -- | P ≤ 0.01 |
| | | AS D. (FF) vs AS E. (FF) | -- | P ≤ 0.01 |

The fifth identified IAPLTR1 proviral insertion occurred in the *LamB3* gene (89). Finally, an IAPLTR1 germline insertion was identified in the *Cdk5rap1* gene (90).

Three cases of IAPLTR1 insertions that meet the above criteria were identified in cell lines: *HoxB8/Hox-2.4*, *Il3*, and *Eps8R1* (91-95). The *HoxB8/Hox-2.4* and *Il3* IAPLTR1 insertions have 100% sequence identity with each other and with the LamB3 insertion in the mouse germline. Since the *Il3* IAP insertion is the best characterized of these cases, this sequence will be referred to as IL3LTR.

To be included on the list of candidates, the IAPLTR1s had to comprise the 5' part of a transcript. Interestingly, all these cases appear to use an antisense promoter within the LTR to drive these transcripts.

All seven sequences together with the four consensus sequences IAPLTR (IAPLTR1 consensus, IAPLTR1_MM consensus, IAPLTR_L1 consensus, and IAPLT_H1 consensus) were aligned and a phylogeny was constructed to 1000 bootstraps in order to see how these sequences
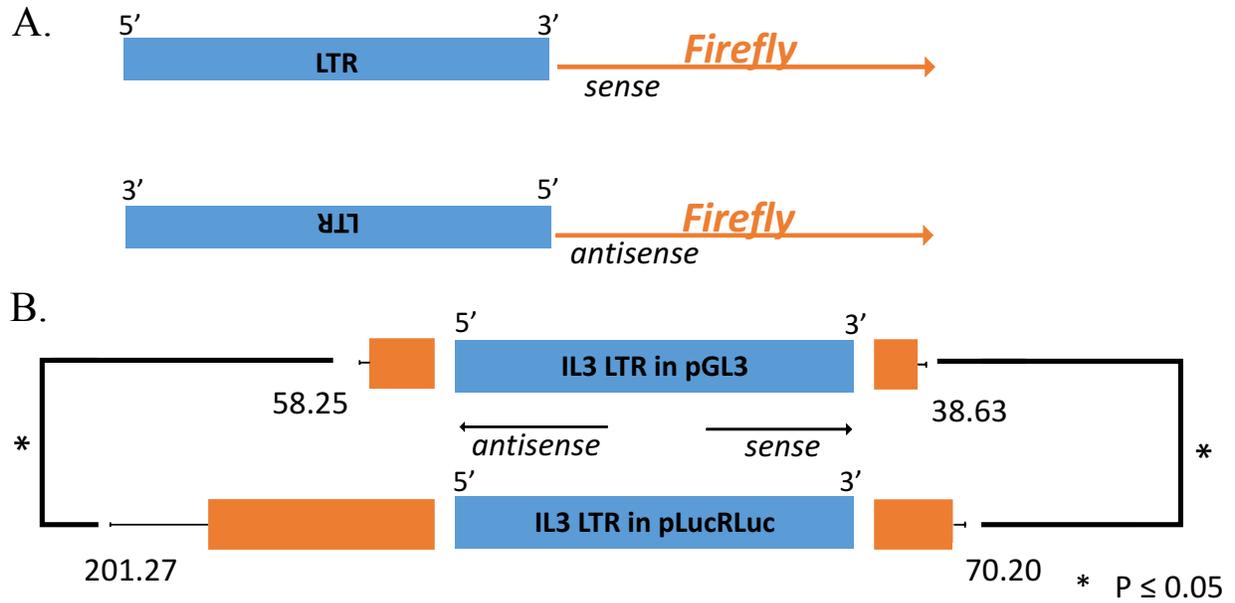
**Figure 3.3  Bidirectional Reporter Constructs Return Different Experimental Results than Unidirectional Reporter Constructs.**

**A. Schematic of an LTR being cloned in both orientations relative to the Firefly luciferase**. Both sense and antisense promoter activity – regardless of whether the experiment was carried out in a unidirectional reporter construct or a bidirectional reporter construct – were measured with Firefly luciferase so that direct comparisons could be made. **B. Comparison of IL3 LTR promoter activity in pGL3 (unidirectional reporter construct) and pLucRLuc (bidirectional reporter construct).** Firefly luciferase activity was found by normalizing over transfection efficiency and finding the normalized fold-value over the empty vector. Error bars are the standard deviation of biological replicates. Blue represents the promoter of interest. Orange is the measurement of Firefly luciferase.

clustered (Figure 3.4) (71). Representative H1 and L1 IAPLTR1 sequences were selected based on how tightly they clustered with the H1 and L1 IAPLTR1 consensus sequences (Figure 3.4).

### 3.2.2    *Wnt9b* IAPLTR1 is representative of the H1 subclass of IAPLTR1

The *Wnt9b* IAPLTR1 insertion had the sequence that clustered most closely to the consensus sequence for the H1 subclass of IAPLTR1 (Figure 3.4). This Wnt9bLTR was chosen as the representative of the H1 subclass of IAPLTR1.

The IAP insertion occurred 6.6kb downstream of the *Wnt9b* gene, in sense orientation relative to the gene (88). A/WySn mice with this IAP insertion and a secondary allele in *clf2* (on another chromosome) present with cleft lip (88). The secondary allele causes incomplete methylation of the Wnt9bLTR, de-repressing the LTR and allowing antisense transcripts to run out of the 5' LTR into the genomic DNA and toward the *Wnt9b* gene (88). These antisense transcripts have an inverse relationship with Wnt9b gene transcripts that coincides with the cleft lip phenotype (88).

### 3.2.3    *Il3* IAPLTR1 is representative of the L1 subclass of IAPLTR1

The IL3LTR clustered closely with the L1 IAPLTR1 subclass consensus sequence; therefore, the IL3LTR was chosen as the L1 IAPLTR1 subclass representative (Figure 3.4). It should be noted that the Rep-base consensus sequence for IAPLTR1 also clusters in this group, indicating that most IAPLTR1 sequences in the genome are probably related to the L1 subclass.

There are at least three identified, independent cases of spontaneous mutation involving sequences identical to the IL3LTR, one in the germline – *LamB3* – and two in the WEHI3b (leukemic) cell line: *Il3* and *HoxB8/Hox-2.4*. The *LamB3* IAPLTR1 insertion occurred in C3H mice (89). A severe blistering disease, junctional epidermolysis bullosa, results from this insertion causing a disruption of the *LamB3* gene (which codes for a laminin-5 subunit) (89). No

transcription has been identified initiating in the IAPLTR1 insertion at *LamB3* (89); however, an

identical LTR forms the 5' end of a transcript in the case of an IAP insertion 215bp upstream of

the *Il3* gene in the Balb/c-derived WEHI3b leukemia cell line (95). This IAPLTR insertion

causes constitutive transcription of the *Il3* gene (95). In the case of the final IAPLTR1 L1

subclass representative insertion, which also occurred in the WEHI3b cell line, the IAP inserted

0.3kb upstream of the *HoxB8/Hox2.4* translation initiation site and induces constitutive

transcription of this gene (92-94).

### 3.2.4    H1 and L1 IAPLTRs are Bidirectional Promoters

To be considered bidirectional, a promoter must generate sense and antisense transcripts

with TSSs less than 1000bp apart and have less than a four-fold bias between initiation of sense

and antisense transcripts (72). Furthermore, bidirectional promoters are typically characterized

by: a specific subset of TFBSs, lack of a clear TATA, a median GC content of around 66%, and

multiple TSSs (*i.e.* 'broad peak' transcription) (72-78). As will be discussed below, H1 and L1

subclasses of IAPLTR1 meet these requirements and should be considered bidirectional

promoters.

### 3.2.4.1    H1 and L1 IAPLTRs Meet the Requirements to be Considered Bidirectional
### Promoters

All IAPLTR1s are from 330-370bp in length – including the R and U5 regions, which

have no known promoter activity (65). The U3 region of IAPLTR1s – the part of the LTR

endowed with promoter capacity – is just over 200bp in length (65). Therefore, IAPLTR1

promoter is clearly small enough, less than 1000bp in length, to be considered a bidirectional

promoter.
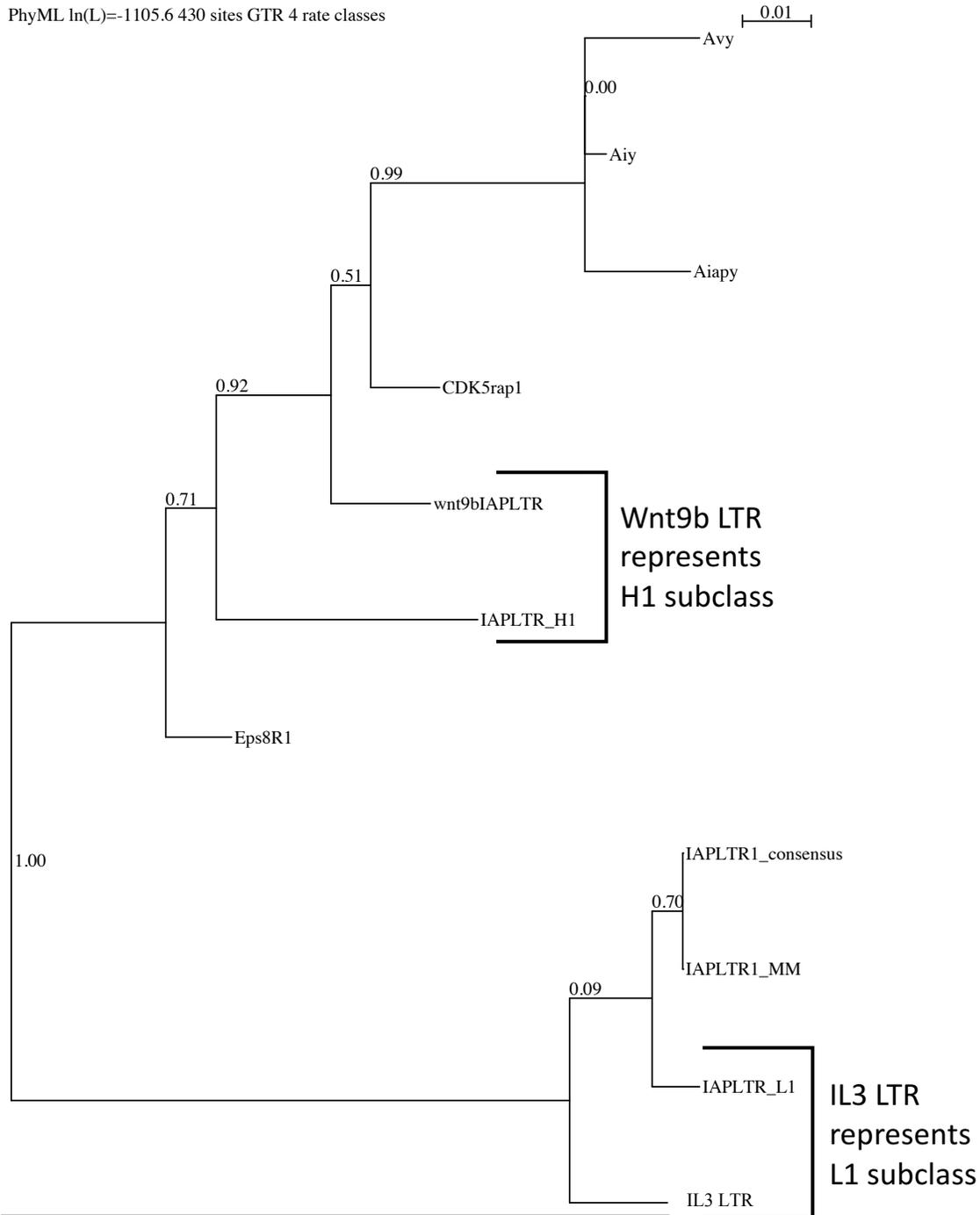
**Figure 3.4 Identification of Functionally Relevant H1 and L1 Subclass Representatives.**

Phylogeny was constructed with the sequences deemed 'functionally relevant' in Section 3.2.1. The sequences that clustered closest to the H1 and L1 consensus sequences (Wnt9bLTR and IL3LTR, respectively) were identified as functionally relevant subclass representatives.

Since it was previously established that Renilla and Firefly luciferases cannot be directly compared, sense activity measured with Renilla was compared to antisense activity also measured with Renilla, and sense activity measured with Firefly was compared to antisense activity measured with Firefly. In all cases – whether measured with Firefly or Renilla luciferase – the H1 and L1 subclasses of IAPLTR1 both reported less than a four-fold difference in sense and antisense promoter activity (Figure 3.2). Thus, H1 and L1 subclasses of IAPLTR1 meet the requirements of a bidirectional promoter.

### 3.2.4.2 H1 and L1 IAPLTR1s Have Many of the Characteristics Typical of a Bidirectional Promoter

Both subclasses of IAPLTR1 also share many of the characteristics typically associated with bidirectional promoters. Bidirectional promoters usually have SP1 motifs, GC-boxes, and an enrichment for BRE and GABP TFBSs. CCAATT-boxes, YY1 motifs, and MYC motifs are overrepresented in bidirectional promoters when compared to unidirectional promoters (List of Abbreviations) (72-78). Except for BRE and GABP, IAPLTR1 sequences typically have each of these motifs (61, 62, 65-69). Wnt9bLTR and IL3LTR also appear to contain these sites – although no Myc site has yet been identified in IAPLTR1s (Figure 3.5) (61, 62, 65-69).

Another characteristic of bidirectional promoters is the lack of a clear TATA sequence. Where 18% of promoters would be expected to contain a TATA by chance, 28% of unidirectional promoters have a TATA sequence – suggesting a selective pressure for TATA – and 8% of bidirectional promoters have a TATA sequence – suggesting a selective pressure against TATA (72). None of the IAPLTR1 sequences examined during the course of this research – including all three consensus sequences for IAPLTR1 – have a clear TATA sequence. Since even small changes to the TATA sequence, like TATTTA to TGAATT, have been

49

demonstrated to cause a significant reduction in promoter activity, the absence of a recognizable TATA sequence strongly indicates that IAPLTR1s are unable to bind the TATA Binding Protein (96).

Bidirectional promoters have a median GC content of 66% (72). IAPLTR1 sequences do not resemble typical bidirectional promoters when it comes to median GC content. The median GC content of IAPLTR1 (measured from the Rep-base consensus sequence) is 50% - less even than the median GC content for unidirectional promoters (which is 53%) (72). Since IAPLTR1s have a lower median GC content than most unidirectional promoters, it is unlikely that the bidirectional promoter activity observed in IAPLTR1s is the result of a high GC content.

Finally, bidirectional promoters typically have broad-peak transcription start sites (TSSs) – in other words, bidirectional promoters usually have multiple TSSs instead of a single, sharp-peaked TSS (which would be typical of a unidirectional promoter). Published 5' RACE experiments demonstrate that IAPLTR1 antisense transcripts have many TSSs (88, 90, 91, 97) (Figure 3.6). IAPLTR1 promoter activity appears to have a broad-peak TSS. The lack of sense TSS analysis should be noted here. Most identified cases in which IAPLTR1s make up the 5' end of a transcript that includes a host sequence are generated from the antisense promoter in IAPLTR1. Thus, there are more reports in the literature of antisense TSSs of IAPLTR1. It should also be noted that many 5' RACE studies only report the longest sequenced RACE-clone, based on the assumption that shorter sequences are artifacts of a fallible assay (example in reference 97). However, when it comes to bidirectional promoter sequences, these studies may have neglected to present evidence for real, alternative TSSs.

All in all, the H1 and L1 subclasses of IAPLTR1 meet the requirements of the definition of a bidirectional promoter and share many of the characteristics that are typically associated

```
                        GC box
                    ┌─────────────┐
                                          Enhancer1
          1      Sp1           Sp1        YY1           GRE                    Z-DNA
wnt9bIAPLTR  TGT-GGGAAG  CCGCCCCCAC  ATTCGCCGTC  ACAAGATGGC  GCTGACATCC  TGTGTTCTAA GTTGGTAAAC  AAATAATCTG  CGCATGAGCC  AAGGGTAT-T
IL3IAPLTR    TGTTGGG-AG  CCGCCCCCAC  ATTCGCCGTT  ACAAGATGGC  GCTGACATCC  TGTGTTCTAA G-TGGTAAAC  AAATAATCTG  CGCATGTGCC  AAGGGTATCT
                          INT                    Sp1                     Enhancer        Sp1
                                                                          Core

          101
wnt9bIAPLTR  TACGACCACT  TGTACTCTGT  TTTTCCCGTG  AACGTCAGCT  CGGCC-ATGG  GCTGCAGCCA  ATCAGGGAGT  GATGCGCGCCT  AGGC-AATGG  TTGTTCTCTT
IL3IAPLTR    TATGACTACT  TGTGCTCTGC  CTTCCCCGTG  -ACGTCAACT  CGGCCGATGG  GCTGCAGCCA  ATCAGGGAGT  GACACGTCCG  AGGCGAAGGA  GAATGCTCCT
                                      AP1                     GC box        CCAAT          Sp1                      TATA

          201     U3  R
wnt9bIAPLTR  TAAAATAGAA  GGGGTTTCGT  TTTTCTCGCT  CTCTTGCTTC  CCTCTCTTGC  TTCTTACACT  CTGGCCCGAT  AAAGATATAA  GCAATAAAGC  TTTGCCGTAG
IL3IAPLTR    TAAGAGGGAC  GGGGTTTTCG  TTTTCTCTCT  CTCTTGCTTC  TTGCTCTCTT  TTCC------  ---------T  GAAGATGTAA  G-AATAAAGC  TTTGCCGCAG
             Txn. Int.                            DPE         DPE                                              poly(A)      R  U5
               Site

          301
wnt9bIAPLTR  AAGATTCTGG  T-TGTTGTGT  TCTTCCTGGC  CGGTCGTGAG  AACGCGTCGA  ATAACA
IL3IAPLTR    AAGATTCTGG  TCTGTGGTGT  TCTTCCTGGC  CGGTCGTGAG  AACGCGTCGA  ATAACA
```

**Figure 3.5 IAPLTR-Associated TFBSs That Appear to be Present in H1 and L1 Subclasses of IAPLTR1.**

These sites were previously identified in IAPLTRs and appear to have intact sites (according to either alignment or TFBS-requirements specified in the literature) in the H1 and L1 IAPLTR1 subclass representatives Wnt9bLTR and IL3LTR, respectively. TFBSs were identified in the following references: 61, 62, 65-69.

Sp1-Specificity Protein 1, INT-integrase site , YY1-Ying Yang 1, GRE-Glucocorticoid Response Element, AP1-Jun/Fos Dimer Binding Site.

51

**Figure 3.6 IAPLTR1s have Many Transcription Start Sites.**

The TSSs represented here were derived from the literature and mapped to their respective TSSs on the above multi-sequence alignment of IAPLTR1 class members. Although there is very little sequence variation between these LTRs, there have been a multitude of antisense TSSs identified for IAPLTR1s. These TSSs w

```
                   301
IAPLTR1_consensus  TCTTGCT--- --CTCTTGCT TCTTGCACTC TGGCTCC-TG AAGATGTAAG CAATAAAGCT TTGCCGCAGA AGATTCTGGT C-TGTGGTGT TCTTCCTGGC
IAPLTR_L1          TCTTGCT--- --CTCTTGCT TCTTGCACTC TGGCTCC-TG AAGATGTAAG AAATAAAGCT TTGCCGCAGA AGATTCTGGT C-TGTGGTGT TCTTCCTGGC
Aiapy              TCTTGCTCT- --CTCTTGCT TCTTGCTCTC TTTTCC--TG AAGATGTAAG -AATAAAGCT TTGTCGCAGA AGATTCTGGT C-TGTGGTGT TCTTCCTGGC
Aiy                TCTTGCTCT- --CTCTTGCT TCTTGCTCTC TTTTCC--TG AAGATGTAAG -AATAAAGCT TTGTCGCAGA AGATTCTGGT C-TGTGGTGT TCTTCCTGGC
Avy                ---------- --CTCTTGCT TCTTGCTCTC TTTTCC--TG AAGATGTAAG -AATAAAGCT TTGTCGCAGA AGATTCTGGT CTTGTGGTGT TCTTCCTGGC
CDK5rap1           TCTTGCGCT- --CTCTTGCT TCTTACACTC TGGCCCGATA AAGATGTAAG -AATAAAGCT TTGTCGCAGA AGATTCTGGT C-TGTGGTGT TCTTCCTGGC
Eps8R1             ----GCTCT- --CTCTTGCT TCTTACGCTC TGGCCCCAT- AAGATGTAAG CAATAAAGCT TTGCCGTAGA AGATTCTGGT --TGTTGTGT TCTTCCTGGC
IAPLTR_H1          -TTGGCGCTG GGCTCTTGCT TCTTACACTC TGGGCTCCCA AAGATGTAAG CAATAAAGCT TTGCCGCAGA AGATTCTGGT C-TGTTGTGT TCTTCCTGGC
wnt9bIAPLTR        -------CT- --CTCTTGCT TCTTACACTC TGGCCCGATA AAGATATAAG CAATAAAGCT TTGCCGTAGA AGATTCTGGT --TGTTGTGT TCTTCCTGGC

                   401
IAPLTR1_consensus  CGGTCGT--- -GAGAACGCG TCTAATAACA
IAPLTR_L1          CGGTCGT--- -GAGAACGCG TCTAATAACA
Aiapy              CGGTCGT--- -GAGAACGCG TCGAATAACA
Aiy                CGGTCGT--- -GAGAACGCG TCGAATAACA
Avy                CGGTCGGTGA AGGAAAGCCG TCGAATAACA
CDK5rap1           CGGTCGT--- -GAGAACGCG ACGAATAACA
Eps8R1             CGGTCGT--- -GAGAACGCG TCGAATAACA
IAPLTR_H1          CGGTCGT--- -GAGAACGCG TCTAATAACA
wnt9bIAPLTR        CGGTCGT--- -GAGAACGCG TCGAATAACA
```

52

with bidirectional promoters. Conclusively then, H1 and L1 IAPLTR1s are bidirectional promoters.

### 3.2.5    L1 is a stronger promoter than H1

To investigate the relative bidirectional promoter activity of the IAPLTR1 subclasses L1 and H1, the L1 IAPLTR1 class representative (IL3LTR) was compared to the H1 IAPLTR1 class representative (Wnt9bLTR), the Firefly activity of L1 was directly compared to the Firefly activity of H1, and the Renilla to Renilla, for each orientation of the LTR. The L1 representative, IL3LTR, was a significantly stronger promoter in sense and antisense orientations than the H1 representative, Wnt9bLTR (Figure 3.2, Table 3.2).

### 3.3    The Core Promoter Sequences of IAPLTR1

The core promoter sequences of IAPLTR1 were investigated next. Core promoters are defined as regions of DNA capable of initiating transcription (98). In this instance, the title of 'core promoter' was conferred to the smallest region of IAPLTR1 that was identified as sufficient to initiate transcription in sense, antisense, or both. It should be noted, however, that although these experiments were accurate enough to identify the core promoter regions, they were not precise enough to identify the minimal core promoter, so an even smaller core promoter could exist within the 'core promoters' identified here.

### 3.3.1    Regions of IAPLTR1 That are Sufficient to Initiate Transcription

In order to identify regions of the LTR that are sufficient to initiate transcription systematic deletions of the L1 subclass were performed, fragments were cloned into the pLucRLuc bidirectional reporter construct, and assayed as above. The L1 subclass of IAPLTR1 was selected for these experiments because it is a significantly stronger promoter than the H1 subclass of IAPLTR1 and therefore provides higher sensitivity for this assay. Primers 1-6, R, and

L were used in conjunction with BestTaq PCR to generate fragments of this LTR named *A-G*, and subsequently cloned into pLucRLuc using XmaI and NcoI cut sites (Figure 3.7). The resulting seven reporter constructs were transfected into p19 cells and, after a 24-hour incubation, the cells were lysed and luciferase activity was measured (Figure 3.7).

Fragment *B* initiated antisense-specific transcription; however, since no transcription was initiated by *A*, the fragment remaining after *A* was subtracted from *B* was identified as a sequence of interest (SOI) (Figure 3.7). This SOI, named SOI1, was expected to contain an antisense-specific core promoter. This region was not sufficient to initiate sense-specific transcription. However, fragments *C*, *D*, and *E* were all capable of initiating sense transcription (Figure 3.7). Since fragments *B* and *F* were insufficient to generate sense transcription, they were subtracted from fragment *D*, and the remaining fragment was hypothesized to contain a core promoter for sense transcription (Figure 3.7). This fragment was named *K*. Fragment *K* also contributes to antisense transcription (Figure 3.7). Its contribution could be that of an enhancer for the antisense promoter identified above or it could contribute a secondary antisense promoter. Therefore, *K* could contain a sense-specific promoter and an antisense-enhancer, or a sense-specific promoter and an antisense-specific promoter, or a core bidirectional promoter.

To investigate these possibilities, the resolution of this experiment was increased. A series of finer, systematic deletions within fragment *K* were performed using primers to generate segments *H-J*. Fragment K was also generated. These segments were cloned into pLucRLuc. The constructs were transfected into p19 cells which were lysed after a 24-hour incubation and luciferase analysis was performed on the resultant lysates.

Fragment *I* was sufficient to initiate antisense-specific transcription (Figure 3.7). However, Fragment *H* was insufficient to generate any significant transcriptional activity, so *H*

was subtracted from *I*, and the remaining fragment was named SOI2 and was expected to contain an antisense-specific core promoter (Figure 3.7).

Fragment *K* was sufficient to initiate sense and antisense transcription (Figure 3.7). Since antisense-specific activity had already been attributed to *I* – *i.e. I* was insufficient to generate significant sense transcription – and *J* was insufficient to initiate any transcription, the fragment remaining after *I* and *J* had been subtracted from *K* was identified as an SOI (Figure 3.7). This SOI, SOI3 – was expected to contain a core promoter for sense transcription. This region could be a bidirectional core promoter, capable of initiating both sense and antisense transcription. However, since the shorter fragment *I* is a stronger antisense promoter than the longer and inclusive fragment *K*, it is more likely that SOI3 contains a repressive element (Figure 3.7). The three SOIs identified here – SOI1, SOI2, and SOI3 – were identified as important, necessary sequences for the transcriptional activity of IAPLTR1. SOI1 and SOI2 appeared to provide a significant contribution to antisense-specific transcription, and were expected to contain antisense-specific core promoters (Figure 3.7). On the other hand, SOI3 seemed to make a significant contribution to sense-specific transcription and was expected to contain a sense-specific core promoter (Figure 3.7).

The SOIs were ordered as oligonucleotide pairs with XmaI and NcoI overhangs. The oligonucleotides were annealed and cloned into pLucRLuc. Each of the resultant SOI-containing pLucRLuc constructs was transfected into p19 cells and, after a 24-hour incubation, the cells were lysed and the lysates analyzed using the luciferase assay, as above.

SOI2 is sufficient to initiate antisense-specific transcription; therefore, this 31bp region was designated an antisense-core promoter for IAPLTR1 (Figure 3.8). However, it should be noted that the antisense promoter activity of this SOI is more than four-times stronger
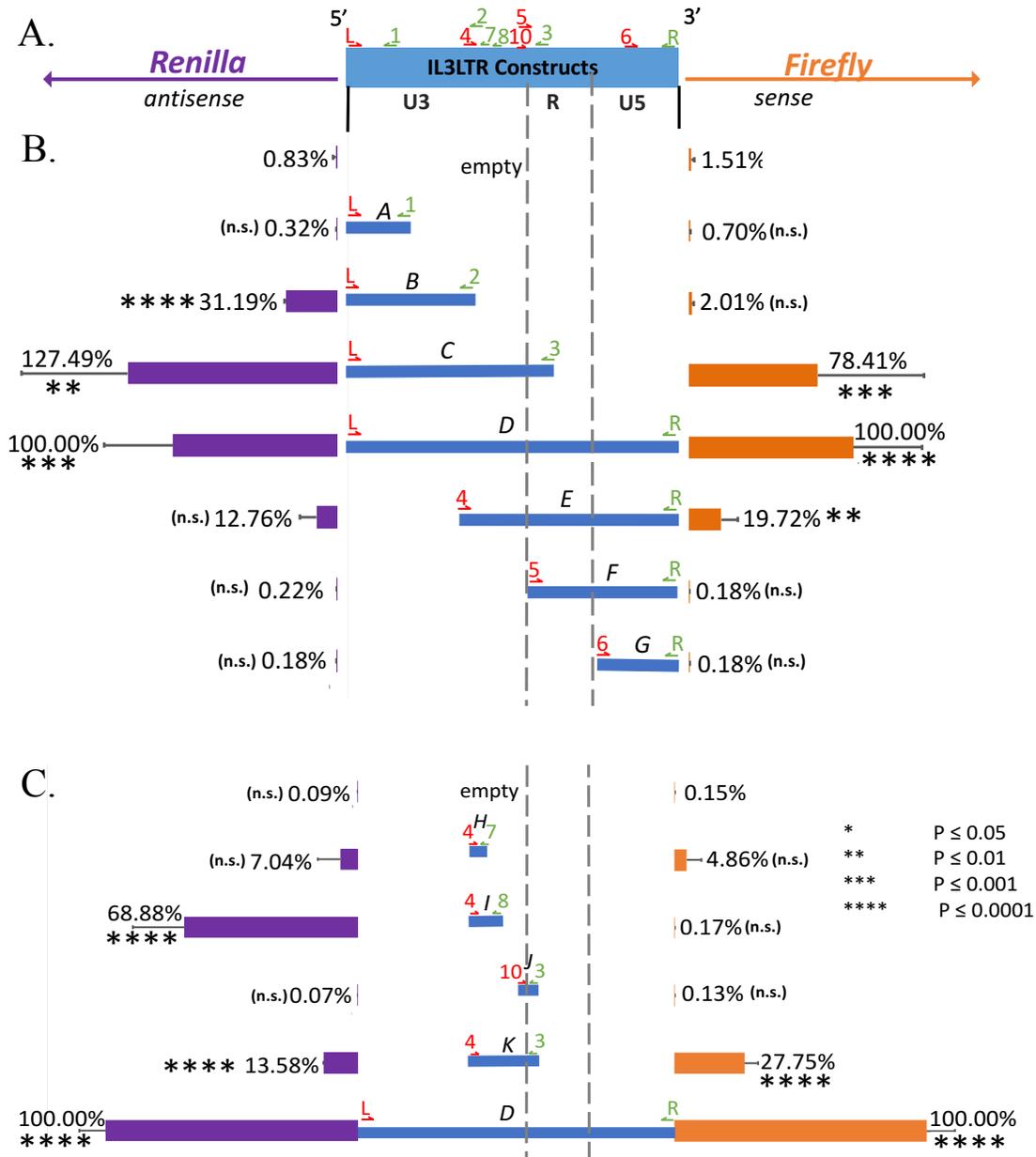
**Figure 3.7 Regions of IL3LTR that are Sufficient to Generate Transcription.**

**(A) Construct Schematic (B) Rough Deletion Assay (C) Fine Deletion Assay.** Primers are indicated in red and green. Fragments names from *A-H* are in black italics. Fragment *B-A* = SOI1 and antisense-specific activity. Fragment *D-B-F* = *K* and bidirectional transcription. Fragment *I-H* = SOI2 and antisense-specific activity. Fragment *K-I-J* = SOI3 and sense-specific activity. Orange is Sense and Firefly. Purple is Antisense and Renilla. Blue is LTR. % is percentage of full-length LTR. (n.s.) is not significant. Significance is measured over the empty vector.

when the 42bp immediately 5' to it are included (*i.e.* fragment *I*) (Figure 3.7). In other words, although SOI2 is sufficient to generate antisense transcription, part or all of fragment *H* is necessary for strong antisense-specific transcription from this core promoter. The added strength of promoter activity conferred by this fragment could be due to the presence of TFBSs that confer enhancer qualities to the core promoter SOI2. Further, SOI2 did not generate any sense transcription, demonstrating that the antisense promoter activity of IAPLTR1 is not simply the result of strong sense promoter activity; antisense-specific transcript initiation has its own core promoter and can be separated from sense promoter activity. Therefore, the presence of antisense promoter activity could be beneficial for IAPs.

Notably, SOI1 is insufficient to initiate any transcription. However, it is part of the larger *B* fragment that confers antisense-specific promoter activity to IAPLTR1 (Figure 3.7, Figure 3.8). Although SOI1 is necessary for this region to initiate transcription, it must require some part of the first 60bp of IAPLTR1 to be sufficient as a promoter of antisense transcripts. Interestingly, this indicates that a second, antisense-specific core promoter occurs in the first 143bp of IAPLTR1 (Figure 3.8). The discovery of two independent antisense-specific core promoters in IAPLTR1 that generate antisense transcripts in the absence of sense transcription provides even more evidence that antisense transcript initiation is important for IAPs.

In contrast, these experiments provided no evidence that sense transcripts can be generated in the absence of antisense transcription in IAPLTR1s. Based on the two previous experimental series, all sense promoter activity was generated from fragment *K* (Figure 3.7). Since SOI2 was identified as a region that only initiated antisense transcripts and fragment *J* was identified as a region that was insufficient to initiate any transcription, SOI3 was identified as a necessary region for sense transcription and the candidate region for the sense-specific core

57

**A.**

5'  2  5  3'
L  1  4 7 8 10 3  6  R

IL3LTR Constructs

Renilla — antisense

Firefly — sense

U3  R  U5

**B.**

L  D  R

**** 100.00%                                                    **** 100.00%

SOI1
** 1.16%                                                        (n.s.) 0.66%

SOI2
**** 15.09%                                                     (n.s.) 0.47%

SOI3
(n.s.) 0.25%                                                    (n.s.) 0.14%

empty
0.30%                                                           0.39%

| | |
|---|---|
| * | P ≤ 0.05 |
| ** | P ≤ 0.01 |
| *** | P ≤ 0.001 |
| **** | P ≤ 0.0001 |

**C.**

Antisense Core Promoter 1

IL3LTR_L →                     ← IL3LTR_D1                          SOI1

TGTTGGGAGCCGCCCCCACATTCGCCGTTACAAGATGGCGCTGACATCCTGTGTTCTAAGTGGTAAACAAATAATCTGCGCATGTGCCAAGGG

Sense Core

Antisense Core Promoter 2

SOI2                          SOI3

TATCTTATGACTACTTGTGCTCTGCCTTCCCCGTGACGTCAACTCGGCCGATGGGCTGCAGCCAATCAGGGAGTGACACGTCCGAGGCGAAGG

IL3LTR_D2 ←                   ← IL3LTR_D7                IL3LTR_D8
← IL3LTR_D4 →

Promoter

U3 | R

IL3LTR_D10 →                  ← IL3LTR_D3

AGAATGCTCCTTAAGAGGGACGGGGTTTTCGTTTTCTCTCTCTCTTGCTTCTTGCTCTCTTTTCCTGAAGATGTAAGAATAAAGCTTTGCCGC

IL3LTR_D5 →

R | U5

AGAAGATTCTGGTCTGTGGTGTTCTTCCTGGCCGGTCGTGAGAACGCGTCGAATAACA

IL3LTR_D6 →                   ← IL3LTR_R

**Figure 3.8 SOIs and Core Promoters.**

**(A) Construct Schematic (B) Transcription from SOIs.** Only SOI2 is sufficient to drive antisense-specific transcription. Primers are indicated in green and red. Orange is Sense and Firefly. Purple is Antisense and Renilla. Blue is LTR. **(C) Core Promoters.** SOI1 was deemed necessary but not sufficient for antisense promoter activity, it requires part or all of fragment A to complete Antisense Core Promoter 1. SOI2 is necessary and sufficient for the Antisense Core Promoter 2. SOI3 was deemed necessary but not sufficient for sense promoter activity. SOI3 requires part or all of fragment J for a sense-specific core promoter, or some or all of SOI2 for a bidirectional core-promoter. Primers are indicated with arrows. Solid lines represent regions that are necessary for core promoter activity. Dashed lines represent regions that are required, at least in part, for core promoter activity. Purple represents antisense specific activity. Orange represents sense specific activity.

promoter (Figure 3.7). Yet, SOI3, although necessary for sense-transcription, was insufficient to initiate transcription in and of itself (Figure 3.8), leaving two options for the region of the IAPLTR1 that confers sense promoter activity: 1. the fragment left after *I* is subtracted from *K* is an independent, sense-specific core promoter; or 2. antisense promoter activity is necessary for sense promoter activity and the entire 136bp *K* fragment (including SOI2) is a bidirectional core promoter from which all IAPLTR1 sense transcription initiates.

In the case of IAPLTR1, at least in the L1 subclass, two independent, antisense-specific core promoters were identified, indicating that antisense transcription may play an important role for IAPs (Figure 3.8). On the other hand, sense transcript initiation from IAPLTR1s requires SOI3 but is conferred either by an independent, sense-specific core promoter including part or all of fragment *J*, or a bidirectional core promoter that overlaps with the antisense-specific core promoter SOI2. Taken together, these observations reveal that antisense promoter activity could be important for these IAPs.

### 3.3.2    Sequence Differences Between the Core Promoters Identified in the L1 Subclass of IAPLTR1 and the H1 Subclass of IAPLTR1 Create Differential Promoter Activity

Sixty-four single nucleotide differences exist between the L1 and H1 subclasses of IAPLTR1 (Figure 3.9). Thirty-three of these differences occur within the U3 region – the region from which all IAPLTR1 transcription is controlled (Figure 3.9). Three single nucleotide variations occur in SOI1, three occur in SOI2, and seven occur in SOI3. Since the SOIs are necessary for transcription initiation (as critical components of core promoters) variation in H1 and L1 sequences should confer different attributes to each subclass of IAPLTR1 (Figure 3.9). The stronger promoter activity of the L1 subclass of IAPLTR1 could be the result of the single nucleotide differences between the SOIs of H1 and L1.

To investigate the contribution of these differences to promoter activity, the three single nucleotide variants present in the SOI1 of H1 were introduced in the L1 subclass representative IL3LTR. Each of these three, single nucleotide differences were introduced individually in the full length IL3LTR in pLucRLuc so that a single change was present in each construct. 24-hour transfections in p19 cells were performed, the cells were lysed, and promoter activity was measured by luciferase analysis. No significant decrease in promoter activity was detected when a single H1-like point mutation was introduced into the L1 IAPLTR1 sequence. However, the individual induction of point mutation three – making a T-to-C change – produced a significant increase in antisense transcription (Figure 3.10).

Every pairing of these three variants was then introduced by point mutagenesis in the IL3LTR and the same experimental assay was performed. A significant decrease in promoter activity was detected with any combination of two Wnt9bLTR-like variants (Figure 3.10). The most significant double point mutation-induced decrease in promoter activity was observed when point mutation two (C-deletion) was combined with point mutation three (T-to-C transition): the overall transcriptional activity was decreased by 85% (Figure 3.10). The least significant double point mutation-induced decrease in promoter activity was observed when point mutation one (T-to-A transition) was combined with point mutation three (T-to-C transition) and the overall transcription was reduced approximately by half (Figure 3.10).

The combinatory effect of all three point mutations in a single sequence had the strongest effect, reducing promoter activity by 95% in sense and in antisense (Figure 3.10). While this represents a drastic reduction in activity, the remaining activity, 5% of the un-modified IL3LTR, was still significantly more than the signal detected from the empty vector. The fact that 95% of total promoter activity can be lost when only 3 point mutations are introduced to the SOI1

A.

```
            1
wnt9bIAPLTR TGT-GGGAAG  CCGCCCCCAC  ATTCGCCGTC  ACAAGATGGC  GCTGACATCC  TGTGTTCTAA GTTGGTAAAC  AAATAATCTG  CGCATGAGCC  AAGGGTAT-T
IL3IAPLTR   TGTTGGG-AG  CCGCCCCCAC  ATTCGCCGTT  ACAAGATGGC  GCTGACATCC  TGTGTTCTAA G-TGGTAAAC  AAATAATCTG  CGCATGTGCC  AAGGGTATCT

            101
wnt9bIAPLTR TACGACCACT  TGTACTCTGT  TTTTCCCGTG  AACGTCAGCT  CGGCC-ATGG  GCTGCAGCCA  ATCAGGGAGT  GATGCGCCCT  AGGC-AATGG  TTGTTCTCTT
IL3IAPLTR   TATGACTACT  TGTGCTCTGC  CTTCCCCGTG  -ACGTCAACT  CGGCCGATGG  GCTGCAGCCA  ATCAGGGAGT  GACACGTCCG  AGGCGAAGGA  GAATGCTCCT

            201                 U3 | R
wnt9bIAPLTR TAAAATAGAA  GGGGTTTCGT  TTTTCTCGCT  CTCTTGCTTC  CCTCTCTTGC  TTCTTACACT  CTGGCCCGAT  AAAGATATAA  GCAATAAAGC  TTTGCCGTAG    R | U3
IL3IAPLTR   TAAGAGGGAC  GGGGTTTTCG  TTTTCTCTCT  CTCTTGCTTC  TTGCTCTCTT  TTCC------  ---------T  GAAGATGTAA  G-AATAAAGC  TTTGCCGCAG

            301
wnt9bIAPLTR AAGATTCTGG  T-TGTTGTGT  TCTTCCTGGC  CGGTCGTGAG  AACGCGTCGA  ATAACA
IL3IAPLTR   AAGATTCTGG  TCTGTGGTGT  TCTTCCTGGC  CGGTCGTGAG  AACGCGTCGA  ATAACA
```

B.

```
SOI1   IL3LTR    TGGTAAAC  AAATAATCTG  CGCATGTGCC  AAGGGTATCT TATGA
       Wnt9bLTR  TGGTAAAC  AAATAATCTG  CGCATGAGCC  AAGGGTAT-T TACGA


SOI2   IL3LTR    GG  GCTGCAGCCA  ATCAGGGAGT  GACACGTCC
       Wnt9bLTR  GG  GCTGCAGCCA  ATCAGGGAGT  GATGCGCCC


SOI3   IL3LTR    G  AGGCGAAGGA  GAATGCTC
       Wnt9bLTR  T  AGGC-AATGG  TTGTTCTC
```

**Figure 3.9 Sequence Variation Between H1 and L1 Subclasses of IAPLTR1.**

**(A) Full LTR Sequence Comparison.** 64 single nucleotide variants exist between Wnt9bLTR and IL3LTR. 33 of these variants occur in the U3 promoter region. **(B) Comparison of SOI Sequences.** SOI1 contains 3 single nucleotide variations. SO2 contains 3 single nucleotide variations. SOI3 contains 7 single nucleotide variations. Sequence variants are highlighted in red. SOIs associated with antisense transcription are coloured purple and the SOI associated with sense transcription is coloured orange.

**Figure 3.10 Wnt9bLTR SOI1 Variants in IL3LTR Reduces Promoter Strength.**

**(A) Construct Schematic (B) SOI1 Variants and Point Mutation Names (C) Single Wnt9bLTR-like SOI1 Point Mutations (D) Combinations of Wnt9bLTR-like SOI1 Point Mutations.** Significant reduction of IL3LTR promoter activity is achieved with any combination of two Wnt9b-like variants introduced by point mutation. All three SOI1 variants reduce IL3LTR promoter activity by 95% reduction in promoter activity is achieved with all three mutations. Orange is Sense and Firefly. Purple is Antisense and Renilla. Blue is LTR.

sequence for IL3LTR suggests these specific nucleotides are critical for proper promoter activity of the L1 subclass of IAPLTR1. It also provides evidence that the few variations between IAPLTR1 subclass representatives are heavily involved in the differential promoter activities of the H1 and L1 IAPLTR1 subclasses. This result also corroborates that the SOIs are, indeed, necessary sequences for the promoter activity of IAPLTR1.

The reciprocal point mutations were then performed, *i.e.* the IL3LTR-specific variants were introduced to the H1 subclass representative, Wnt9bLTR, in the pLucRLuc construct and the assays were repeated. The opposite experiment was expected to produce the opposite effect: instead of a reduction in promoter activity, the introduction of the stronger L1 IAPLTR1 promoter variants into the weaker H1 IAPLTR1 promoter sequence was expected to increase promoter activity. In this case, only three constructs were made: a single point mutation, two point mutations, and all three point mutations.

The single (C-insertion) and double (C-insertion + C-to-T transition) point mutations produced no significant effect on the antisense activity of the LTR, and, instead of increasing transcriptional activity, reduced the sense transcription by half (Figure 3.11). This unexpected result indicates that there may be a repressive element for sense transcription present at the 3' edge of SOI1 in the L1 IAPLTR1. Although the first point mutation in Wnt9bLTR was the addition of a cytosine residue, the second point mutation – which, if anything, produced stronger sense-repression than the first point mutation – removed a cytosine residue, indicating that the repressive effect cannot be due to an increase in methylated cytosine residues. If a repressive sequence is present, it must interact with a transcription factor or induce a particular conformation of DNA. Alternatively, the introduction of these point mutations could cause the loss of a necessary sense-specific TFBS. This, however, is unlikely, given that these two bases in

the otherwise identical L1 SOI1 seem to be responsible for 85% of the sense transcriptional activity of the IL3LTR (Figure 3.10).

Even more intriguing was the combinatory effect of all three L1 LTR-like point mutations in the H1 IAPLTR: the presence of all three L1 variants in SOI1 of Wnt9bLTR, completely restored the sense promoter activity that had been lost with the first and second point mutations, and removed antisense promoter activity entirely (Figure 3.11). Collectively, these three point mutations cause a loss of bidirectional promoter activity, providing the first evidence that IAPLTR1 promoters have the ability to promote sense transcription without antisense transcription. Although there is no evidence that sense-specific promoter activity is possible in the naturally existing IAPLTR1 sequence, leaving open the possibility that IAPLTR1s transitioned to an antisense-dominant promoter due to the presence of or lack of a selective pressure. Together, these point mutation experiments demonstrate the complexity of IAPLTR promoters.

The few differences between these highly similar IAPLTR1 subclass sequences appear to be of functional relevance, as evidenced by the drastic transcriptional differences introduced by introducing the three point mutations in L1 SOI1 so that it resembles H1 SOI1 and vice versa. This level of functional relevance fits well with the model of rapidly evolving, highly conserved TFBSs within LTRs (41). Indeed, if the functional differences between the L1 and H1 IAPLTR1s are due to the presence and absence of TFBS, then it should be possible to identify TFBSs present in L1 that are not present in the same regions of H1 and vice versa.
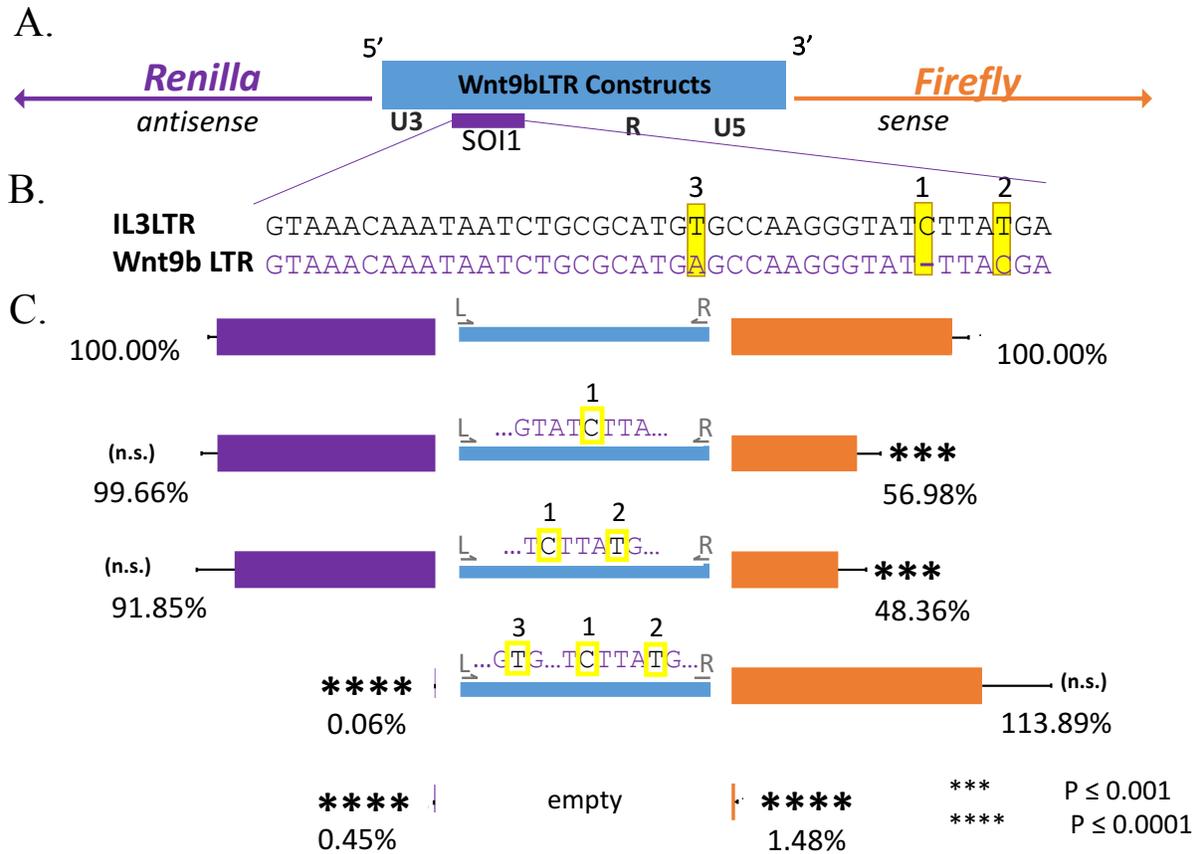
**Figure 3.11 IL3LTR SOI1 Variants in Wnt9bLTR Remove Antisense Promoter Activity.**

**(A) Construct Schematic (B) SOI1 sequence Variants and Point Mutation Numbers (C) Combinatory Effect of All Three IL3-like Variants Causes Loss of Antisense Transcription.** The first IL3-like variant introduced in Wnt9bLTR caused a 50% reduction in sense promoter activity. This reduction was maintained with the second mutation, but regained with the third where it was also accompanied by the complete loss of antisense promoter activity. Orange is Sense and Firefly. Purple is Antisense and Renilla. Blue is LTR.

**3.4    Variation in the Transcription Factor Binding Sites of the U3 Region of H1 and L1**

**IAPLTR1 Subclasses Coincide with Differential Promoter Activity**

**3.4.1.1    TFBS Variation and Differential Promoter Activity of H1 and L1 Subclass**

**Representatives.**

Common TFBSs – TFBSs occurring in the same location of the sequences being

compared – could be critical TFBSs for general promoter function (since they appear to be

conserved between sequences), but they are unlikely to contribute significantly to the differential

promoter activity in H1 and L1 IAPLTR1 subclasses. As described below, although there are

only 33 single nucleotide variants between the U3 regions (~220bp enhancer-promoter regions)

of H1 and L1 IAPLTR1 subclass representatives, there were 32 unique TFBSs predicted (Figure

3.12). These unique, putative sites could be responsible for the differences in promoter activity

of H1 and L1 subclasses.

In the U3 region of the naturally occurring H1 and L1 representative sequences 32

unique, putative TFBSs were discovered using the TFBS-finding software oPPOSUM3.0 and

PROMO-ALGGEN (99, 100). Fourteen of these unique sites were found in the L1 representative

IL3LTR. Eleven of the IL3LTR binding sites were for known activators (NOBOX, TFE3-S,

MYB, MYC, NKX2-5(x2), c-JUN(x2), HIFA, REL, and CREB1), and two were for TFs that

could activate or repress transcription (MZF1 and KLF4) (101). No known repressive TFBSs

were predicted uniquely for IL3LTR. Perhaps the lack of a unique repressive sequence and the

presence of four unique activation TFBSs with TFs present in the early embryo (therefore in p19

cells) is what endows IL3LTR with stronger promoter activity than Wnt9bLTR.

Eighteen unique, putative TFBSs were identified in Wnt9bLTR, the H1 representative.

Eight activator TFBSs were identified in this sequence (c-JUN(x2), AP1, FOXA2, FOXO3,

NFATC2, SPIB, NKX2-5), five TFBSs that could activate or repress transcription (YY1, MAZ, GATA1, MZF1(x2)), and five that repress transcription (NKX3-2, MIZF(x2), CTCF, ZNF354c) (101). Wnt9bLTR had unique TFBSs predicted for repressors and activators that are present in the early embryo (likely present in p19 cells); these repressive sites could contribute to Wnt9bLTRs weaker promoter activity in p19 cells.

### 3.4.2   TFBS Variation and IAPLTR1 SOIs

Three SOIs were identified as necessary regions for promoter activity of IAPLTR1 (Figure 3.7). These regions were identified using IL3LTR, the L1 IAPLTR1 subclass representative. There are very few variations between IL3LTR and Wnt9bLTR (the H1 IAPLTR1 subclass representative) (Figure 3.9); however, these variants result in a number of different predicted TFBSs present at the SOIs (Figure 3.12).

Earlier, SOI1 was determined to be necessary but not sufficient to generate antisense activity from one of the two antisense core promoters identified in IAPLTR1 (Figure 3.8). Three intact, putative TFBSs were discovered in this region: sites for activating TFs MYC, TFE3-S, and NKX2-5 (101) (Figure 3.13). Partial sites were also predicted on either edge of SOI: sites for activator NOBOX and repressor NKX3-2 on the 5' end of the SOI, and a site for activator c-JUN on the 3' end of the SOI (Figure 3.13). Perhaps the inclusion of complete sites for the TFBSs on either edge of SOI1 would make this region sufficient for promoter activity. This SOI region of the Wnt9bLTR has a putative site for activating TF FOXA2 interrupted at the 5' end; an intact, activating TFBS (AP1) present in the same location as IL3LTRs MYC and TFE3-S sites; and an interrupted 3' TFBS for repressor/activator MAX and repressors ZNF354c and NKX3-2 (101) (Figure 3.13). The presence of a TFBS for pluripotency factor MYC in the IL3LTR, and the
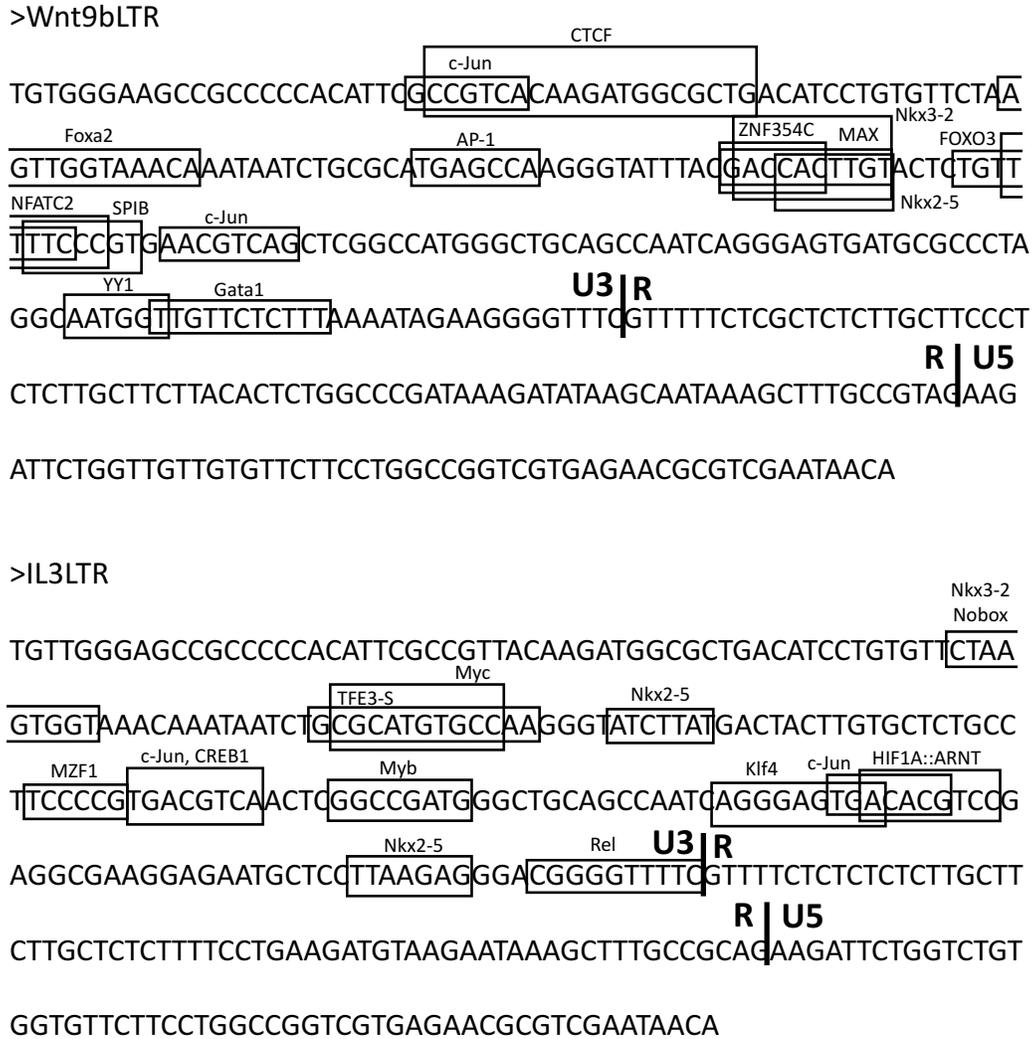
>Wnt9bLTR

CTCF
c-Jun
TGTGGGAAGCCGCCCCCACATTCGCCGTCACAAGATGGCGCTGACATCCTGTGTTCTAA

Foxa2        AP-1         ZNF354C  MAX  Nkx3-2  FOXO3
GTTGGTAAACAAATAATCTGCGCATGAGCCAAGGGTATTTACGACCACTTGTACTCTGTT

NFATC2  SPIB  c-Jun                                    Nkx2-5
TTTCCGTGAACGTCAGCTCGGCCATGGGCTGCAGCCAATCAGGGAGTGATGCGCCCTA

YY1   Gata1                              U3|R
GGCAATGGTTGTTCTCTTTAAAATAGAAGGGGTTTCGTTTTTCTCGCTCTCTTGCTTCCCT

                                        R|U5
CTCTTGCTTCTTACACTCTGGCCCGATAAAGATATAAGCAATAAAGCTTTGCCGTACAAG

ATTCTGGTTGTTGTGTTCTTCCTGGCCGGTCGTGAGAACGCGTCGAATAACA


>IL3LTR

                                                Nkx3-2
                                                Nobox
TGTTGGGAGCCGCCCCCACATTCGCCGTTACAAGATGGCGCTGACATCCTGTGTTCTAA

              TFE3-S  Myc        Nkx2-5
GTGGTAAACAAATAATCTGCGCATGTGCCAAGGGTATCTTATGACTACTTGTGCTCTGCC

MZF1  c-Jun, CREB1   Myb           Klf4  c-Jun  HIF1A::ARNT
TTCCCCGTGACGTCAACTCGGCCGATGGGCTGCAGCCAATCAGGGAGTGACACGTCCG

              Nkx2-5        Rel      U3|R
AGGCGAAGGAGAATGCTCCTTAAGAGGGACGGGGTTTTCGTTTTCTCTCTCTCTTGCTT

                              R|U5
CTTGCTCTCTTTTCCTGAAGATGTAAGAATAAAGCTTTGCCGCAGAAGATTCTGGTCTGT

GGTGTTCTTCCTGGCCGGTCGTGAGAACGCGTCGAATAACA

**Figure 3.12 Uniquely Predicted TFBSs for H1 and L1 Subclass Representatives.**

These are the putative TFBSs that were discovered using oPPOSUM3.0 and PROMO-AGGLEN that were unique to either the IL3LTR sequence or the Wnt9bLTR sequence.

absence of such a site in Wnt9bLTR, could play a role in the stronger promoter activity observed for IL3LTRs in pluripotent p19 cells.

SOI2 was identified as the second antisense core promoter present in IAPLTR1 – both necessary and sufficient to generate antisense transcription from the second antisense core promoter in IL3LTR (Figure 3.8). This region contains three intact, putative TFBSs in IL3LTR – activators c-JUN and HIF1A, and activator/repressor KLF4 – as well as an interrupted, putative TFBS for activating TF MYB at the 5' end of this SOI (Figure 3.13) (101). There were no unique TFBSs present in H1 IAPLTR1 in this SOI region; thus, each of these activators could be contributing to the stronger promoter activity of IL3LTR (Figure 3.13). It would be interesting to test the SOI2 region of Wnt9bLTR to see if it is sufficient to produce antisense activity without any of these unique TFBSs. If not, these TFBSs could be responsible for one of the antisense core promoters in IL3LTR.

SOI3 was identified as a region that was necessary but not sufficient for sense transcription in IAPLTR1s (Figure 3.7, Figure 3.8). Interestingly, in IL3LTR – the LTR that was used for the promoter bashing experiments – no unique TFBSs were predicted (Figure 3.13). Conversely, two unique TFBSs were predicted in Wnt9bLTR: activator/repressor TFs YY1 and GATA1 (GATA1 is an interrupted TFBS, it overlaps the 3' edge of the SOI3 region in Wnt9bLTR) (Figure 3.13) (101). Perhaps this region shares a sense-specific activator TFBS. However, of the SOIs, this region has, by far, the most variation per nucleotide so it is less likely to share TFBSs (Figure 3.9). It is possible that sense transcription may originate from a bidirectional core promoter that includes the second antisense-core promoter (SOI2).

Altogether, the variation between TFBSs in H1 and L1 subclass representatives suggests that the identified SOIs may not be SOIs in both subclasses. The SOIs identified are likely for the
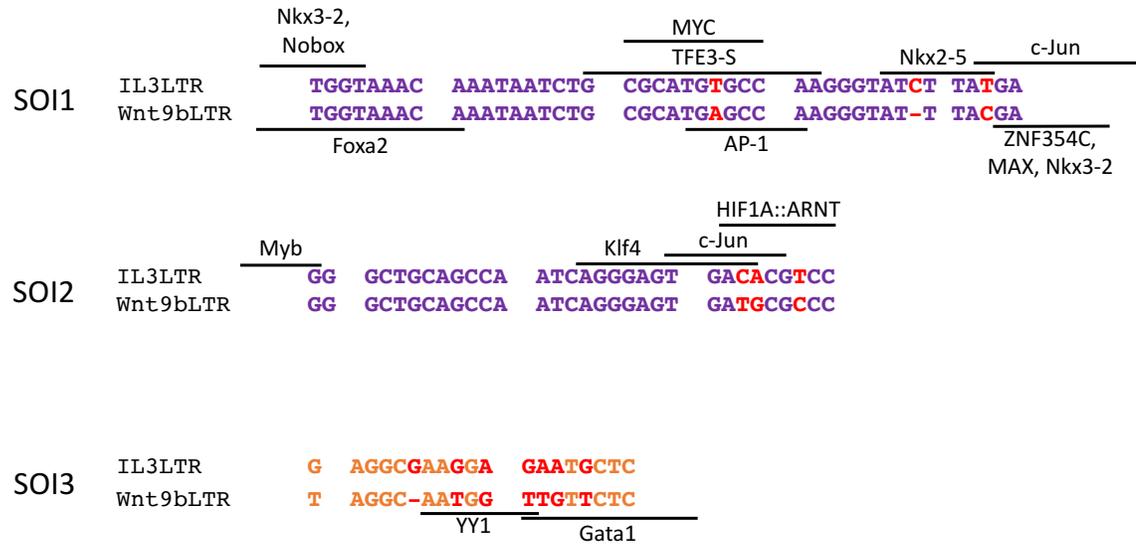
**Figure 3.13 Uniquely Predicted TFBSs in the SOIs for H1 and L1 Subclasses of IAPLTR1.**

Black lines indicate where the putative TFBS lies or overlaps neighboring sequence. TFBSs above are for IL3LTR and below are for Wnt9bLTR. Purple is for antisense specificity. Orange is for sense specificity. Red is for variants.

L1 subclass specifically. A second set of experiments would be required to see if the SOIs of the H1 subclass are the same regions of the LTR, even though the TFBSs are different.

### 3.4.3    TFBS Variation and SOI1 Point Mutation Experiments

The Wnt9bLTR-like point mutations that were introduced in SOI1 of IL3LTR changed the TFBSs predicted in that SOI (Figure 3.14). Point mutation one caused a loss of MYC and TFE3-S putative binding sites, as well as a gain of AP1 (Figure 3.14). Point mutation two caused a loss of the NKX2-5 putative binding site and a gain of FOXI1 (Figure 3.14). Point mutation three caused a loss of NKX2-5 and c-Jun putative binding sites (Figure 3.14). Of these three point mutations, only point mutation three produced a significant change in promoter activity – it resulted in a significant increase in antisense promoter activity. The other two point mutations, when introduced separately, did not significantly change the promoter activity of IAPLTR1

70

(Figure 3.10). However, any and all combinations of these point mutations (and the subsequent loss/gain of TFBSs) resulted in a significant decrease in both sense and antisense promoter activity (Figure 3.10). Loss of MYC, TFE3-S, and NKX2-5 putative TFBSs coincides with reduced promoter activity, suggesting that the combination of these three sites is important for the promoter function of IL3LTR in p19 cells (Figure 3.14). This reduced promoter capacity may be exacerbated by the loss of the putative c-Jun site at the 3' end of the IL3LTRs SOI1; however, these results are inconclusive with respect to the effect of the presence of a c-Jun motif in SOI1 (Figure 3.14). The gain of the AP1 putative binding site cannot compensate for the MYC, TFE3-S, NKX2-5 triple-binding site loss, although it may be able to compensate for the loss of MYC and TFE3-S in the presence of NKX2-5 (Figure 3.14). The gain of a FOXI1 site does not appear to have a significant effect on the promoter activity of the IL3LTR in p19 cells (Figure 3.14).

The IL3LTR-like point mutations that were introduced in the SOI1 of Wnt9bLTR also changed the TFBSs predicted in the SOI1 (Figure 3.14). Point mutation one did not cause any change in the TFBSs identified in the SOI1 region of Wnt9bLTR; however, it did cause a significant loss of sense promoter activity. The addition of point mutation two created a putative c-Jun site at the 3' end of SOI1, but did not change the promoter activity from the state induced by point mutation three. Therefore, the acquisition of this c-JUN TFBS does not appear to play a significant role in promoter capacity of Wnt9bLTR. The addition of the third point mutation caused a loss of the putative AP1 site and the gain of a MYC site and a TFE3-S site. The acquisition of MYC and TFE3-S putative binding sites accompanied by the loss of AP1, either in concert with or independent from the acquired putative c-JUN site downstream, resulted in a complete loss of antisense promoter activity and full restoration of sense promoter activity (Figure 3.14). In the SOI1 region of Wnt9bLTR, the presence of MYC and TFE3-S – two

71

predicted, activator TFBSs – appear to coincide with the loss of antisense transcription. It is

possible that the predicted AP1 site in Wnt9bLTR is important for antisense transcript initiation.

### 3.4.4    TFBS Variation and Sensitivity to Ectopic Transcription Factor Expression

Since the SOIs are apparently necessary (if not always sufficient) for core promoter

activity, TFBSs that are predicted in an SOI in H1 or L1 but not both IAPLTR1 subclass

representatives are good candidates for TFBSs that contribute to differences in promoter activity.

If these differentially predicted TFBSs are important to IAPLTR1 function, then the subclass

representative with the TFBS may be sensitive to changes in the amount of that TF present in the

cell type under study, while the subclass representative without the TFBS should be insensitive

to the amount of that TF. Five transcriptional activators that were predicted to have binding sites

in an SOI region of IL3LTR – the subclass representative with stronger promoter activity – were

selected as candidates for contribution to the stronger promoter activity of the L1 subclass of

IAPLTR1. MYC and TFE3-S were selected because they were two of three differential TFBSs

that were identified as important to transcription initiation in IL3LTR (Figure 3.14). c-JUN was

selected because it was predicted in two SOIs and because it was predicted in the cluster of three

differential TFBSs predicted in SOI2, the only SOI that was both necessary and sufficient for

promoter activity (Figure 3.14). These first three TFs were also selected because they all had

putative binding sites affected by the point mutation experiments in SOI1 – the predicted sites

were lost with all three point mutations in IL3LTR and gained with all three point mutations in

Wnt9bLTR. KLF4 and HIFA were the final two TFs selected because they were the other two

differentially predicted TFBSs in SOI2 (Figure 3.13).

Published ectopic expression constructs for each of these TFs were ordered from

Addgene and co-transfected into p19 cells with pLucRLuc containing either IL3LTR or
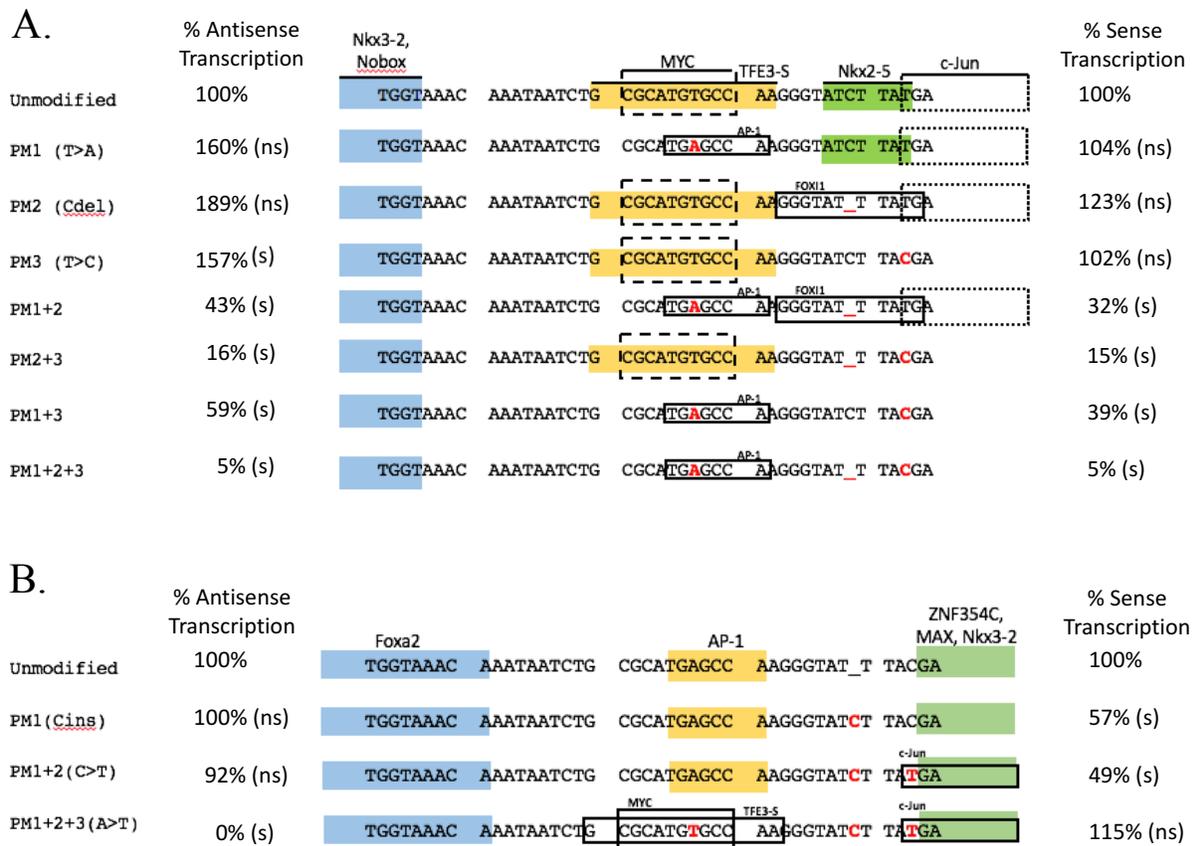
**Figure 3.14 SOI1 Point Mutation Experiments and Predicted TFBSs.**

**(A) Wnt9bLTR-like Variants Introduced in the SOI1 of IL3LTR. (B) IL3-like Variants Introduced in the SOI1 of Wnt9bLTR.** Predicted TFBSs are boxed with solid lines, colour, or dashed lines. Red is for variants introduced by point mutation. Percentage of sense and antisense transcription relative to the unmodified sequence are listed on either side of the SOI1 sequence. (s) is significantly different from 100% expression. (ns) is not significantly different from 100% expression.

Wnt9bLTR and pCMV-Bgal (102-105). The IL3LTR was expected to respond to every case of ectopic expression of a TF with a significant increase in promoter activity relative to Wnt9bLTR (which was expected to be relatively insensitive to the ectopic expression of the TF). This result was not obtained in any of the ectopic expression assays (Figure 3.15). In all cases, there was no significant difference between the response of IL3LTR in the presence of the ectopically expressed TF/IL3LTR under normal conditions compared to the response of the Wnt9bLTR in the presence of ectopically expressed TF/Wnt9bLTR under control conditions (Figure 3.15).

KLF4 and MYC are pluripotency factors, so they are already present – likely in abundance – in p19 cells (which is pluripotent teratocarcinoma cell line) (101). It is possible that ectopic expression of these TFs did not significantly affect transcription because they were already highly expressed in these cells. It is also possible that ectopic expression of a TF in these cells promoted the initiation of differentiation. Differentiation could explain the consistency of the variation in promoter activity between IL3LTR and Wnt9bLTR (106). Sensitivity to the presence/absence of TFs with differential binding sites in IL3LTR and Wnt9bLTR may be better assayed by knocking down TFs of interest with RNAi or by performing transient transfections in cells that have the TF of interest knocked out.
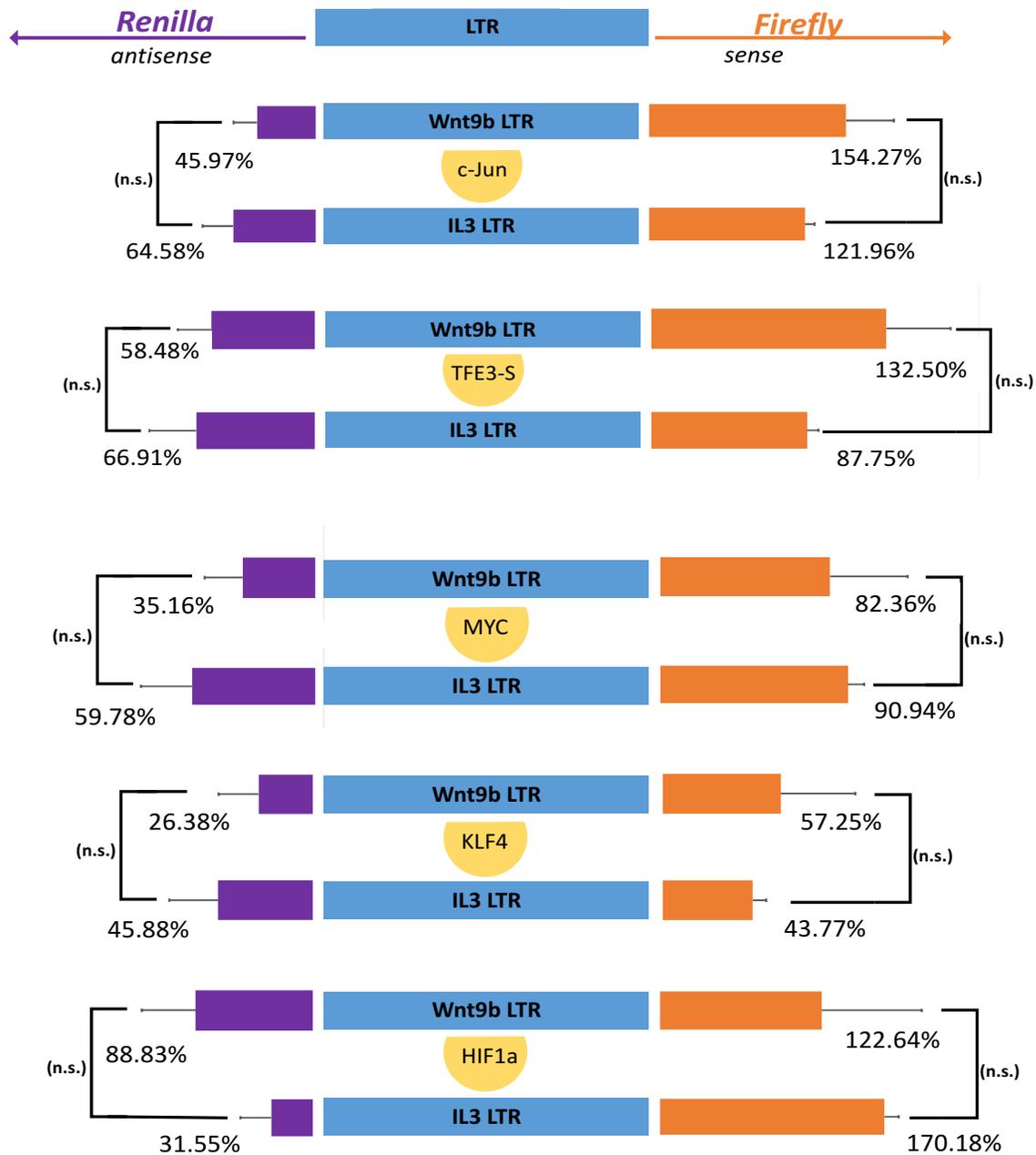
**Figure 3.15 TF Overexpression Assays.**

No significant difference in sensitivity to the overexpression of these TFs of interest was found between IL3LTR and Wnt9bLTR. Orange is Firefly and Sense. Purple is Renilla and Antisense. Blue is LTR. Yellow indicates the ectopically expressed transcription factor. (n.s.) indicates 'not significant'. % is promoter activity in the presence of the overexpressed TF over promoter activity in the absence of the overexpressed TF.

## Chapter 4: Discussion and Conclusions

There is a growing interest in bidirectional promoters. Recently, several bidirectional reporter vectors have been constructed and published in studies in which researchers attempted to understand bidirectional promoters and their activity. This research provided the first in-depth analysis of bidirectional reporter constructs and documented the comparison of these constructs to traditional, unidirectional reporter construct for analysis of bidirectional promoters. The findings were conclusive: bidirectional reporter vectors reduce error and remove bias, and, therefore, are better for the analysis of bidirectional promoters.

IAPLTRs have long been known to have antisense promoter activity. These experiments provide evidence that IAPLTR1s – a young and abundant subclass of IAPLTRs – are bidirectional promoters (59). They meet the requirements of bidirectional promoters and have many characteristics that typify bidirectional promoters. In this research, functionally relevant representatives were identified for the sequence-based high (H1) and low (L1) divergence subclasses that Qin *et al.* suggested for IAPLTR1 (59). A detailed study of these sequences with the use of a bidirectional reporter vector revealed that their divergence has functional implications and results in differential promoter performance. Putative TFBSs unique to these differences in sequence were predicted *in silico* and five candidate TFs were ectopically expressed in the presence of H1 and L1 subclass representatives, but no significant difference in sensitivity to the ectopic expression of the TF was observed. Finally, regions of IAPLTR1 that contain the core promoter sequences were identified. Two independent regions were sufficient to promote antisense-specific activity; only one region was sufficient to promote sense activity and no evidence was found to suggest that this sense activity could be driven independently of antisense activity.

## 4.1    IAPLTR1 Core Promoter Sequences

A seminal study in 1988 tested the bidirectional promoter capacity of four IAPLTR sequences (65). Although there were no established criteria with which to define a bidirectional promoter, this paper demonstrated that IAPLTR1a sequences could initiate sense and antisense transcription – a trait that is not shared among all IAPLTRs (65). There was significant variability in promoter activity within this class of elements (65). Using a unidirectional reporter construct, regions important for sense and antisense transcription in IAPLTR1a were identified (65).

The data presented here show that IAPLTR1 can also initiate transcripts in sense and antisense, and can be defined as bidirectional promoter according to the current criteria (72-78). Using a bidirectional reporter construct, two antisense-specific core promoter regions were identified, and one region was identified as important for sense transcription. Antisense Core Promoter One and the region important for sense promoter activity lie within the regions that were previously identified as important for antisense and sense transcription in the closely related IAPLTR1a class of elements (Figure 4.1) (65). However, the Antisense Core Promoter Two was previously unidentified as important for antisense transcription in any IAPLTR. These experiments also report significant variability in promoter activity for IAPLTRs within the same class.

## 4.2    Importance of Antisense Promoters to LTR Activity

The presence of two, independent regions that are sufficient to drive antisense-specific transcription suggests that antisense transcription is important for IAPLTR1s. To date, there have been two well-defined cases of antisense transcription from an LTR conferring a selective advantage to its retroviral element (107). Exogenous retroviruses HTLV-1 and HIV-1 both have

antisense LTR activity that drives transcription 3'-to-5' into the viral sequence from the 3' LTR (107). HTLV-1 has an antisense open reading frame (ORF) that generates the HBZ protein which induces/maintains HTLV-1 latency and causes proliferation of the infected cell (107). HIV-1 has an antisense transcript that induces/maintains HIV-1 latency and an antisense ORF that produces the ASP protein that modulates autophagy (107). It is possible that the redundancy of antisense promoters in the IAPLTR1 sequence is present to ensure the production of similar, yet undiscovered, beneficial antisense transcript or ORF.

Alternatively, the presence of antisense transcription from the 5' LTR could benefit the IAP simply by promoting a local open chromatin state (82, 83). This would provide a rationale for the apparent importance of antisense transcription – *i.e.* why there appears to be redundant antisense-specific promoters, but no clear sense-specific promoter. If strong antisense transcription from the LTR promotes open chromatin and relatively few sense transcripts are required for successful retotransposition, it would even be logical if sense transcription was the result of strong antisense transcription. Most IAPLTR1s are associated with internal deletions, and so are non-autonomous, requiring full length element proteins (59). So, in general, they are probably not responsible for producing full-length element proteins. Although high sense-transcript copy number could give a stoichiometric advantage to finding and binding the necessary full-length element proteins, there are other ways of out-competing other non-autonomous transcripts for those proteins – such as increased binding efficiency, coordinated transcription, *etc*... Further, strong antisense transcription could reduce the number of IAP transcripts and fewer IAP transcripts could be easier for the cell to tolerate. This is especially important to note given the importance of antisense transcription to exogenous retroviruses in entering and maintaining the proviral state (107).

## 4.3 Putative IAPLTR1 Subclass-Specific TFBSs, IAPLTR-Associated TFBSs, and Bidirectional Promoter-Associated TFBSs

In this research, putative TFBSs unique to each of the H1 and L1 IAPLTR1 subclass representatives were identified using TFBS-finding software. These sites were represented (in red) alongside TFBSs that have previously been identified in IAPLTRs (in black) (Figure 4.1). Except for the second GC-box, which was only present in the IL3LTR, and the fifth SP1 site, which was only present in the Wnt9bLTR, all TFBSs that had been identified in the literature were present in both IAPLTR1 sequences and tended to lie in regions that did not accumulate sequence differences (Figure 4.1) (61, 62, 65-69.). It is possible that these sites tend to be conserved in IAPLTRs.

Interestingly, many of the IAPLTR-associated TFBSs are also typically found in bidirectional promoters (indicated with a blue star) (Figure 4.1). GC-boxes, SP1 sites, YY1 sites, and a CCAAT box are all common both to bidirectional promoters and IAPLTRs (72-78). The absence of a recognizable TATA sequence in both IAPLTR1s is not common to all IAPLTR sequences (some IAPLTRs have recognizable TATA sequences), but is a typical feature of bidirectional promoters that is common to many IAPLTRs (72).

Bidirectional promoter-associated or IAPLTR-associated putative TFBSs were also identified in regions of the IAPLTR1 that were deemed necessary for transcription (Figure 3.7). SOI1 (Sequences of Interest were identified as regions of the IAPLTR1 sequence that are necessary for promoter activity) of both IAPLTR1 subclass representatives overlaps the Enhancer1/Enhancer Core that is typically found in IAPLTRs and contains an SP1 site – a motif common to bidirectional promoters (Figure 4.1). The L1 representative, IL3LTR, also has a unique, putative motif for the bidirectional promoter-associated, transcriptional activator MYC
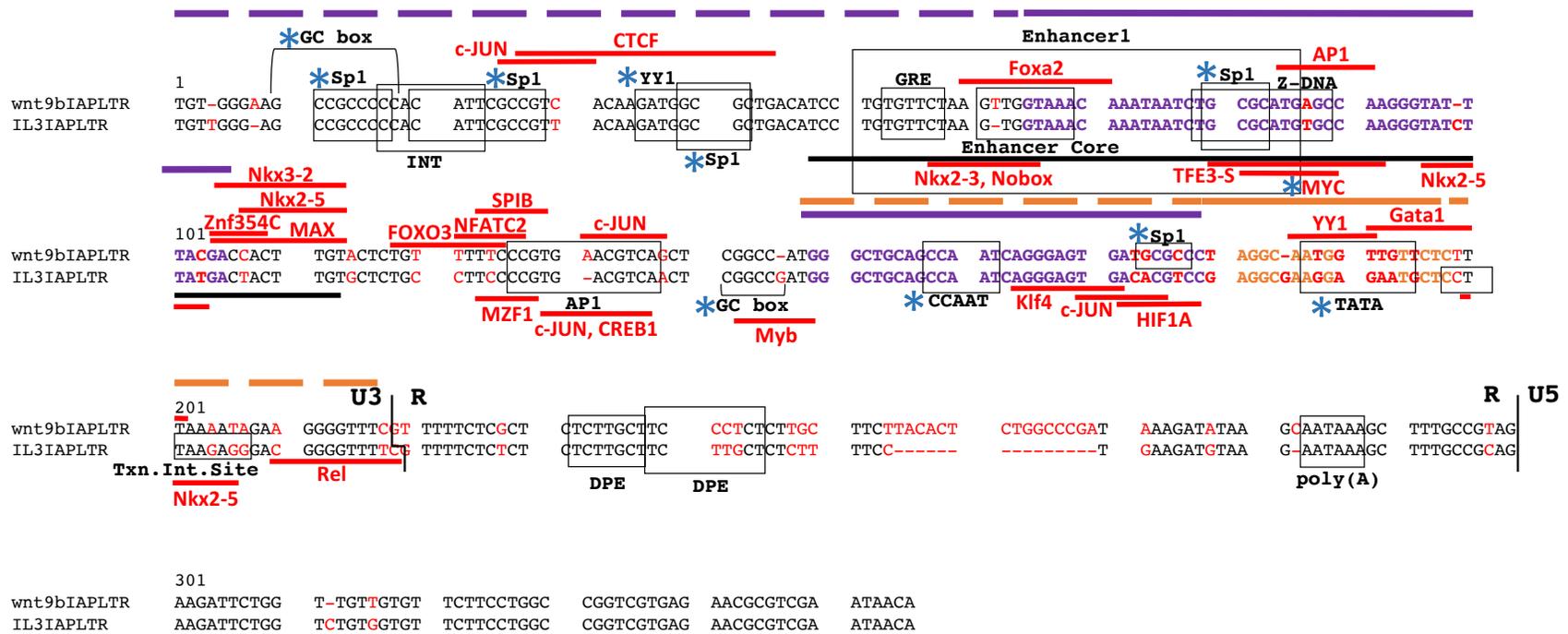
**Figure 4.1 Important Regions of IAPLTR1 and Putative Unique TFBSs.**

Here all IAP-associated TFBSs present in the H1 and L1 Subclass Representatives of IAPLTR1 are represented in black alongside the unique TFBSs predicted for each subclass in red (Wnt9bLTR-specific=above; IL3LTR-specific=below). Blue stars indicate TFBSs that are also characteristically present in bidirectional promoters. Antisense and sense core promoter sequences are represented by purple and orange lines above the sequence, respectively (dashed line represents region that is required at least in part for sufficiency of the necessary solid-line sequence; where no dashed line is present, sequence is necessary and sufficient for promoter activity). Sequence variants are red, antisense-specific SOIs are purple, sense-specific SOIs are orange. Definitions here: List of Abbreviations.

[Some data from references: 61, 62, 65-69.]

(Figure 4.1). SOI2 contains the CCAAT box – commonly found in both IAPLTR and

bidirectional promoter sequences – and another putative SP1 site in the H1 representative (Figure

4.1). SOI3 has the disrupted TATA sequence in both IL3LTR and Wnt9b LTR; however, the

transcription initiation site common to IAPLTRs is only intact in the IL3LTR (Figure 4.1).

Altogether, the bidirectional promoter associated TFBSs predicted in the IAPLTR1 subclass

representatives were typically common sites to both H1 and L1 IAPLTR1s and to IAPLTR

sequences in general. This may be an artifact of a literature bias, since most IAPLTR studies

have been carried out on a subset of IAPLTR1a sequences (a class that is closely related to

IAPLTR1).

## 4.4    H1 and L1 Subclass Nomenclature Should be Adopted for the IAPLTR1 Class of ERVs

H1 and L1 subclasses of IAPLTR1 were suggested based on divergence (high (H1) and

low (L1)) from the Rep-base consensus sequence. The representative sequences for the H1 and

L1 subclasses were examined to find out whether the sequence divergence was accompanied by

a functional divergence.

Although there are only thirty-three single nucleotide differences between the U3 regions

(the promoters) of the H1 and L1 subclass representatives selected, there were many putative,

unique TFBSs associated with the sequences of each. Presumably, one or more of the TFs that

was predicted to bind a subclass-specific sequence variant contributes to the significantly

stronger promoter ability of the L1 IAPLTR1 subclass over the H1 IAPLTR1 subclass.

Regardless of whether the TF(s) responsible for this observed difference were successfully

identified, the point mutation assays provide definitive evidence that the sequence divergence

between H1 and L1 IAPLTR1 subclasses is accompanied by a functional divergence (Figure 3.10, Figure 3.11).

When the SOI1 of the L1 subclass representative of IAPLTR1 was replaced in the L1 sequence with the SOI1 from the H1 subclass representative (by point mutation for the three single nucleotide variations), 95% of total promoter activity – in sense and in antisense – was lost (Figure 3.10). In the reciprocal experiment, where the SOI1 of the H1 subclass representative was replaced in the H1 IAPLTR1 sequence with the SOI1 from the L1 representative, the H1 promoter ceased to have bidirectional activity: 100% of sense promoter activity was maintained but all antisense activity was lost (Figure 3.11). The SOI1 sequences of the H1 and L1 representatives have only three single nucleotide variations between them (Figure 3.9). The striking impact that these three variants have on their respective sequences reveals that there is a divergence in function that accompanies the H1 and L1 divergence in sequence from the IAPLTR1 consensus. Altogether, these results provide functional evidence that H1 and L1 subclass nomenclature should be applied to the IAPLTR1 class of elements.

## 4.5   IAPLTR1 Expansions and ERV-Host Co-Evolution

When the sequence divergence of all identified IAPLTR1 sequences in the mouse genome is plotted against the number of sequences with 'x' millidivs of divergence, there appear to be three expansions of IAPLTR1 sequence, which I refer to as X, Y, and Z (Figure 4.2) (59). Although it is tempting to use the Rep-base consensus sequence for IAPLTR1 as the most similar sequence to the ancient exogenous retrovirus (aXRV), the aXRV and the expansion(s) most closely related to it remain unknown. However, Y and Z are apparently more closely related to each other than to X based on phylogenetic clustering (59). This allows for the construction of three general options for the evolution of IAPLTR1: 1. Y and Z share an
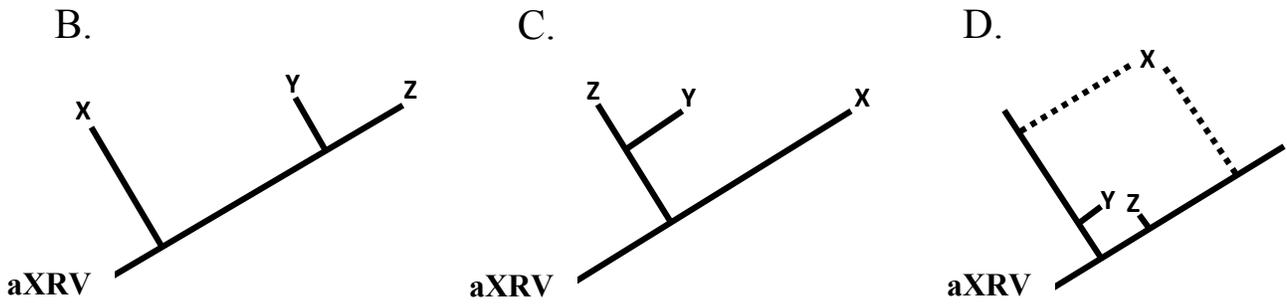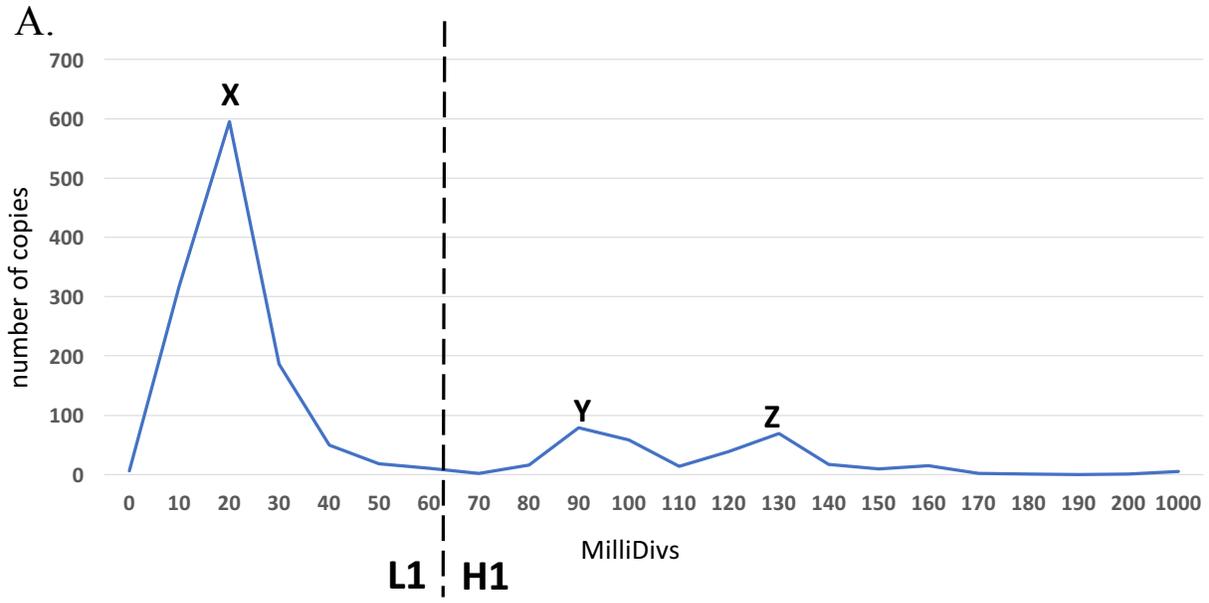
**Figure 4.2 IAPLTR1 Expansions and Evolution.**

**(A) Three Expansions** of IAPLTR1 appeared to have occurred, based on sequence divergence. These expansions were named left-to-right X, Y, and Z. MilliDivs are 1bp divergence for every 1000bp (from RepBase consensus sequence for IAPLTR1) [Data from reference 59] **(B-D) Evolutionary Models for IAPLTR1 Expansions**. Based on phylogenetic analysis [in reference 59], Z and Y sequences are more closely related to each other than either of them is to X. Option B shows X as the most closely related sequence to aXRV. Options C and D show Z and Y as the most closely related sequences to aXRV: Option C represents a scenario in which Z and Y share the same LCA with X; alternatively, Option D represents two scenarios (each indicated by a dotted line to X) in which either Y or Z shares a more recent LCA with X than the other – in this case, Y and Z are still more closely related to each other than they are to X.

**Table 4.1 Comparison of Cell Type and Background for 'Active' IAPLTR1 Subclasses.**

| IAPLTR1 Subclass | Insertion Name (Genbank #) | Transcription or Retrotranposition Activity | Cell Type for Activity | Background | Reference |
|---|---|---|---|---|---|
| L1 | Fused | Spontaneous Mutation | Germ Line | C57BL6 | 11 |
| L1 | IL3 (X04120) | Drives Constitutive Expression of IL3 | WEHI3b cells (leukemic cell line) | Balb/c | 95 |
| L1 | HOX2.4 (X54077) | Drives Constitutive Expression of Hox2.4 | WEHI3b cells (leukemic cell line) | Balb/c | 93 |
| L1 | c-mos (AH001881.2) | Activates c-mos | Plamacytoma | Balb/c | 108 |
| L1 | Notch4 (AB016771) | Generates an alternatively spliced Notch4 transcript | Mammary Tumour | Balb/c | 109 |
| L1 | Knobbly | Spontaneous Mutation | Germ Line | C3H | 11 |
| L1 | IL6 (X51457) | Drives Constitutive Expression of IL6 | Plasmacytoma | Balb/c | 110 |
| L1 | GM-CSF-959 | Drives Constitutive Expression of CSFs | FDC-P1 Cells (leukemogenesis) | DBA/2 | 111 |
| L1 | GM-CSF-180 | | | | |
| H1 | Agouti (L33247) | Spontaneous Mutation | Germ Line | C57BL6 | 86 |
| H1 | ProtolAP (AB099818) | Common Retrotransposition in Leukemia | Meyloid Leukemia | C3H | 112 |
| H1 | FBP1 (M97701) | Increases Expression of FBP1 | Leukemic Cell Line (L1210) | DBA | 113 |
| H1 | IL3 (D63766) | Retrotranposition Event | Meyloid Leukemia (L-8028) | C3H/He | 114 |
| H1 | Ep (AF004352) | Retrotranspotition Event | Germ Line | C3H/FeJ | 115 |
| H1 | Ap3d1 (AH012499.2) | Retrotransposition Event | Germ Line | C3H/HeJ | 116 |
| H1 | A$^{iapy}$ | Spontaneous Mutation | Germ Line | C3H/HeJ | 86 |
| H1 | A$^{vy}$ | Spontaneous Mutation | Germ Line | C3H/HeJ | 85 |
| H1 | A$^{iy}$ | Spontaneous Mutation | Germ Line | C3H/HeJ | 87 |
| H1 | Eps8R1 | Drives Ectopic Expression of Eps8R1 | Many Tissues | C57BL6 | 91 |
| H1 | Cdk5rap1 (BB842254) | Polymorphic Insertion Provides poly(A) | Many Tissues | C57BL6 | 90 |

ancestral sequence that is more closely related to the aXRV for IAPLTR1 than X; 2. Y and Z share an ancestral sequence that is more distantly related to the aXRV for IAPLTR1 than X; 3. a variation of option one in which X shares an ancestral sequence with either Y or Z, but is more distantly related to Y or Z than they are to each other (Figure 4.2).

A comparison of 'active' (published cases of IAPLTR1s either driving transcription or recently retrotransposed) H1 and L1 IAPLTR1s, as identified and subcategorized by this study, or by the research of Qin *et al.* (59), uncovered trends for IAPLTR1 activity (summarized in Table 4.1). Under normal, developmental conditions, members of the H1 subclass of IAPLTR1 appear more likely to retrotranspose than their L1 counterparts (Table 4.1). Spontaneous germline mutations involving IAPLTR1of either subclass occur most often in C3H mice (Table 4.1).

Qin *et al.* had previously asserted that L1 IAPLTRs were more likely to be expressed in blood cancers as transforming factors than in normal blood cells – whereas H1 IAPLTRs could be expressed in normal blood cells without acting as transforming factors (59). The stronger promoter activity observed for the L1 subclass of IAPLTR1 may be the reason for the L1 subclass' increased likelihood of driving transforming, ectopic expression of endogenous genes. Perhaps the tolerance blood cells show for H1-initiated transcription could be attributed to the weaker promoter activity observed for the H1 subclass of IAPLTR1. Notably, while the C3H strain of mouse appears to have more germline retrotransposition events of IAPLTR1s than any other, the Balb/c strain appears to harbour more IAPLTR1-related neoplastic transformations than any other (Table 4.1). Balb/s'c tumour susceptibility could be related to its immune function, given that it has a proclivity for developing inflammatory diseases (117).

TFBSs for immune/stress response factors made up 50% of all unique TFBSs predicted for both subclasses, suggesting that IAPLTR1s could be responsive to the host's stress and immune response (Table 4.2). Involvement of ERVs in immune/stress response has previously been identified as a mode through which ERVs can confer benefits to the host (27, 52-56). This could also explain why, of all possible cancers, IAPLTR1s seem to participate in blood-cell derived cancers.

ERVs have also been implicated in driving the formation of and reshaping the epigenome (17, 19, 33). Where only 16.7% of putative, unique TFBSs identified in L1 relate to epigenomic factors and relate to the formation of open chromatin, 57.1% of factors with putative binding motifs identified in H1 had to do with the epigenome and the majority were associated with the formation of closed chromatin – including a KRAB-zinc finger protein (ZNF354C) (Table 4.2). The lack of L1-specific epigenetic repressive motifs alongside the presence of H1-specific epigenetic repressive motifs combined with the fact that there are far more known cases of H1 activity under normal conditions, provides some evidence that the X expansion (L1) was an earlier expansion than Z and Y (H1) and that it is usually repressed by factors that H1 has developed modes of escaping.

Most of the putative, unique TFBSs discovered in both IAPLTR1s were for TFs that are expressed in the early embryo, in embryonic reproductive structures, and/or in meiotic oocytes – in other words, most predicted, subclass-specific TFs are present when they need to be if they do indeed affect the ability of the IAP to colonize the genome (Table 4.2). It should be noted that both H1 and L1 subclasses of IAPLTR1 independently acquired predicted sites for NKX3-2, NKX2-5, and – most interestingly – c-JUN. c-JUN is of particular interest because, of the

uniquely predicted TFBSs, it is the only one with a TF that has known expression in primordial oocytes (Table 4.2).

Furthermore, the H1 subclass of IAPLTR1 has a greater diversity of putative unique binding sites between its consensus sequence and its representative sequence (Wnt9bLTR) in spite of the fact that it has fewer total sequences than the L1 subclass (Figure 4.2, Table 4.2) (59). This divergence is probably due to the two individual expansions that have taken place and the ensuing directional selection for particular sequences. The divergence in this subclass may confer higher replicative fitness to the H1 subclass of IAPLTR1 over the L1 subclass, because it has fewer elements that will be responsive to a broader range of environmental stimuli. Even in these short sequences, it appears that increased variation within a replication-capable population results in increased fitness.

**Table 4.2 Summary of TFs Predicted to Bind Unique Sequences in IAPLTR1 Subclass Representatives and Consensus Sequences.**

[This data is from references 101 and 118-119] GS=genitourinary system; M=meiotic; (P)=primordial; UE=unfertilized egg; ICM=inner cell mass; RS=reproductive structures; BL=blastocyst; PF=pluripotency factor. Note that E10.5 is when the P germ cells migrate to the reproductive structures.

| IAPLTR1 Name (Subclass) | Name of TF | Activator/ Repressor | Orientation Relative to LTR | Unique Sites | Embryonic Expression | Oocyte Expression (M/P) | Sperm Expression | Stress Response Gene? | Involvement in Epigenome |
|---|---|---|---|---|---|---|---|---|---|
| Wnt9b (H1) | ZNF354C | Repressor | Sense | 1 | E3.5 E10.5(GS) | No (M) | No Data | NO. | Recruits histone methyl-transferases |
| | CTCF | Repressor | Sense | 1 | E0.5-E13.5 E15.5 (RS) | Yes (M) | No Data | YES. Anti-apoptotic; cytokine stress response. | Forms chromatin loops |
| | c-JUN | Activator | Antisense | 2 | E3.0-P (BL, ICM, RS) | Yes (P&M) | Yes (P) | YES. Widespread. | Reverses methylation of DNA |
| | AP1 | Activator | Sense | 1 | | | | | |
| | FOXA2 | Activator | Antisense | 1 | E5.5-P | No Data | No Data | YES. Response to acute liver injury. | Interacts with histones to open chromatin |
| | FOXO3 | Activator | Antisense | 1 | E0.5-E3.0 (ICM) E14.5-15.5 (RS) | No (M) Yes (Primary & Secondary Oocytes) | No Data | YES. Hypoxia response; dendritic cell modulation. | None |
| | NFATC2 | Activator | Sense | 1 | E15.5 (RS) | No Data | No Data | YES. T cell activation. | Induces H3K4me3 in dendritic cells |
| | MIZF/ HINFP | Repressor | Antisense/ Sense | 2 | E0.5-E14.5 (BL, ICM) | Yes (M) | No Data | NO. | Interacts with MBD2 |
| | MZF1 | Activator or Repressor | Antisense/ Sense | 2 | No | No Data | No Data | NO. | None |

88

| IAPLTR1 Name (Subclass) | Name of TF | Activator/ Repressor | Orientation Relative to LTR | Unique Sites | Embryonic Expression | Oocyte Expression (M/P) | Sperm Expression | Stress Response Gene? | Involvement in Epigenome |
|---|---|---|---|---|---|---|---|---|---|
| | GATA1 | Activator or Repressor | Antisense | 1 | E1.5-13.5 (BL) | Unclear | No Data | YES. Platelets; erythroid proliferation. | None |
| Wnt9b (H1) | MAX | Activator or Repressor | Antisense | 1 | E1.5-P (BL, ICM, RS) | No Data | No Data | NO. | H3K9 methyl-transferase recruiter |
| | SPIB | Activator | Antisense | 1 | Not Enough Data | No Data | No Data | YES. Immune cell-proliferation and function. | None |
| | YY1 | Activator or Repressor | Antisense | 1 | E1.5-E14.0 (BL, ICM) E16.0-P.0 (RS) | No Data | No Data | NO. | Recruits histone de-acytlases and histone acetyl-transferases |
| | NKX2-5 | Activator | Antisense | 1 | No | No | No | NO. | None |
| | NKX3-2 | Repressor | Antisense | 1 | E14.5 (RS) | No Data | No Data | NO. | None |
| IL3 (L1) | KLF4 (PF) | Activator or Repressor | Antisense | 1 | E1.5-E13.5 (BL, ICM) | Yes (M) | No Data | YES. Heat response; hypoxia. | None |
| | NKX3-2 | Repressor | Sense | 1 | E14.5 (RS) | No Data | No Data | NO. | None |
| | NOBOX | Activator | Sense | 1 | No | Yes (M) | No Data | NO. | None |
| | TFE3-S | Activator | Sense | 1 | E1.5-E13.5 (BL, ICM) E15.5 (RS) | Yes (M) | No Data | YES. Innate Immune Response | None |
| | MYB | Activator | Sense | 1 | E14.5 (RS) | No Data | No Data | NO. | None |
| | MYC (PF) | Activator | Sense | 1 | E2.0-E12.0 (BL, ICM) | Yes | Yes | NO. | Recruits hist. acetyl-transferases |

| IAPLTR1 Name (Subclass) | Name of TF | Activator/ Repressor | Orientation Relative to LTR | Unique Sites | Embryonic Expression | Oocyte Expression (M/P) | Sperm Expression | Stress Response Gene? | Involvement in Epigenome |
|---|---|---|---|---|---|---|---|---|---|
| IL3 (L1) | NKX2-5 | Activator | Antisense/ Sense | 2 | No | No | No | NO. | None |
| | c-JUN | Activator | Sense | 2 | E3.0-P (BL, ICM, RS) | Yes (P&M) | Yes (P) | YES. Widespread. | Reverses methylation of DNA |
| IL3 (L1) | HIF1A | Activator | Antisense | 1 | E1.5-P (BL, ICM) | Yes (M) | No Data | YES. Hypoxia | None |
| | REL | Activator | Sense | 1 | E13.5 (RS) | No Data | No Data | YES. Immune response. | None |
| | MZF1 | Activator or Repressor | Antisense | 1 | No | No Data | No Data | NO. | None |
| | CREB1 | Activator | Antisense | 1 | Not Enough Data | No Data | No Data | YES. Immune response; anti-apoptotic; maintenance of T-cells | None |
| IL3 with PM | FOXI1 | Activator | Sense | 1 | Not Enough Data | No Data | No Data | NO. | None |
| H1 Consensus Only | FOXD3 | Repressor (PF) | Sense | 1 | E6.5-E14.0 E14.5(RS) | No Data | No Data | NO. | None |
| | ELF5 | Activator | Sense | 1 | E6.5 | No Data | No Data | NO. | None |
| | FEV | Repressor | Antisense | 1 | No Data | No Data | No Data | NO. | None |
| | EGR1 | Activator | Antisense | 1 | E1.5 E4.5(ICM) E10.5 (GS) E13.5 (RS) | Yes (M) | No Data | YES. Tumor suppressor; involved in myeloid differentiation | None |
| | | | | | | | | | |

| IAPLTR1 Name (Subclass) | Name of TF | Activator/ Repressor | Orientation Relative to LTR | Unique Sites | Embryonic Expression | Oocyte Expression (M/P) | Sperm Expression | Stress Response Gene? | Involvement in Epigenome |
|---|---|---|---|---|---|---|---|---|---|
| H1 Consensus Only | SOX5 | Activator | Antisense | 1 | E14.5-E15.5 (RS) | No Data | No Data | YES. Th17 differentiation and immune response | None |
| | TBP | Activator | Sense | 1 | E0.5-E13.0 (BL, ICM, RS) | Yes (M, UE) | No Data | NO. | None |
| IL3 Consensus Only | KLF4 (PF) | Activator or Repressor | Sense | 1 | E1.5-E13.5 (BL, ICM) | Yes (M) | No Data | YES. Heat response; hypoxia. | None |

## 4.6    Future Directions

Although I identified regions that contain the core promoters for the H1 and L1 subclasses of IAPLTR1, the minimum core promoter sequences have yet to be elucidated. These sequences could be identified by continuing with fine, 5bp extensions and/or reductions to the 5' and 3' ends of each of the SOIs and by analyzing expression from the LTR construct in pLucRLuc transiently transfected into p19 cells (similarly to the protocols I used in this research).

Using TF prediction software, I identified 32 unique, putative TFBSs that bound the U3 region of either H1 or L1 IAPLTR1 subclass representatives, yet whether these TFs are important for the promoter activity of IAPLTR1s is yet to be determined. The TFs that bind these sequences could be knocked down with siRNA in p19 cells and transient transfections of H1 or L1 IAPLTR1-pLucRLuc constructs could be performed and analyzed as was done here. Some TFs of interest cannot be knocked down in p19 cells, *i.e.* pluripotency factors cannot be knocked down in a pluripotent cell type without causing differentiation which would alter the expression of IAPLTR-driven transcripts. Knock downs in p19 cells would also be difficult because they divide rapidly, so a different and slower-growing cell type may be necessary for this experiment. Alternatively, a cell line with a knock-out for a TF of interest could simply have the appropriate constructs transiently transfected into them for assessment. Either the knock out or the knock down of a TF of interest would reveal whether that TF is involved in the promoter activity of that particular subclass of IAPLTR1; however, further experiments would be required to tease out the direct effects from the absence of the TF from the indirect effects.

I demonstrated that bidirectional promoter activity is generated from both subclasses of IAPLTR1 and that the L1 subclass of IAPLTR1 is a stronger promoter than the H1 subclass of IAPLTR in p19 cells. However, it is not yet known whether this promoter activity is affected by

cell type/environment. The same transfection experiments could be performed in alternate cell types, like NIH3T3 (mouse fibroblast) cells. This would provide the opportunity to assay the IAPLTR promoters in a different cellular environment and analyze whether their promoter activity changes in a different (in the case of NIH3T3 cells, differentiated) environment.

I showed that there is strong antisense promoter activity from both subclasses of IAPLTR1; however, it is not yet known whether the 3' LTR of IAPLTR1 elements promotes transcription of an antisense IAP transcript. To discover an antisense transcript for IAPs with IAPLTR1s, the internal region of the IAP with only the 3' LTR (no 5' LTR) could be cloned in sense into a unidirectional reporter construct like pGL4. These constructs could then be transiently transfected into non-murine cells (*i.e.* cells that do not have any background IAP transcripts, like Human K562 cells) and lysate analysis could be performed to determine whether the IAPLTR is transcriptionally active (the 3' LTR should be sufficient to drive sense transcription into the reporter). If a non-murine cell type that is capable of driving transcription from IAPLTR1 can be found, then those cell lysates could be further analyzed with strand-specific RT-PCR and primer walking. This should identify any antisense IAP transcripts generated by the antisense IAPLTR1 activity.

I noted that putative TFBSs unique to either the H1 or L1 subclass of IAPLTR1 tend to have TFs expressed in the early embryo or the stress response, or have TFs that affect the epigenomic state. Global analysis of putative, unique TFBSs among related ERVs from different expansions could provide insight as to whether there is a tendency to acquire certain types of TFBSs. For example, TFBSs that would be responsive to stress response or TFBSs that allow for epigenome modifications, *etc*… This kind of global analysis, especially if paired with functional assays, could uncover trends in the co-evolution of the host and the ERV.

93

## 4.7    Significance of Work

In this work, I identified core promoter sequences of IAPLTR1, provided evidence that the small sequence divergence between two IAPLTR1 expansions was accompanied by functional divergence, and recommended that a system of nomenclature that had previously been suggested based on the aforementioned sequence divergence be accepted. I also noted that IAPLTRs may have a tendency to acquire stress-response and epigenome-related TFBSs, and commented on the nature of host-ERV co-evolution. I discussed three evolutionary expansions of IAPLTR1, and proposed three hypotheses for the evolutionary history of these expansions. Finally, I performed an in-depth analysis of the advantages of using a bidirectional reporter vector over a unidirectional reporter vector when assaying bidirectional promoters.

# References

1. Lander, E., Int Human Genome Sequencing Consortium, Linton, L., Birren, B., Nusbaum, C., Zody, M., . . . Int Human Genome Sequencing Conso. (2001). Initial sequencing and analysis of the human genome. *Nature, 409*(6822), 860-921. doi:10.1038/35057062
2. Waterston, R., Lindblad-Toh, K., Birney, E., Rogers, J., Abril, J., Agarwal, P., . . . Mouse Genome Sequencing Consor. (2002). Initial sequencing and comparative analysis of the mouse genome. *Nature, 420*(6915), 520-562. doi:10.1038/nature01262
3. Mager, D. L., & Stoye, J. P. (2015). Mammalian endogenous retroviruses. *Microbiology Spectrum, 3*(1), MDNA3-0009-2014. doi:10.1128/microbiolspec.MDNA3-0009-2014
4. Boehne, A., Brunet, F., Galiana-Arnoux, D., Schultheis, C., & Volff, J. (2008). Transposable elements as drivers of genomic and biological diversity in vertebrates. *Chromosome Research, 16*(1), 203-215. doi:10.1007/s10577-007-1202-6
5. Mandal, P. K., & Kazazian, H. H., Jr. (2008). SnapShot: Vertebrate transposons. *Cell, 135*(1), 192.e1. doi:10.1016/j.cell.2008.09.028
6. Magiorkinis, G., Gifford, R. J., Katzourakis, A., De Ranter, J., & Belshaw, R. (2012). Env-less endogenous retroviruses are genomic superspreaders. *Proceedings of the National Academy of Sciences of the United States of America, 109*(19), 7385-7390. doi:10.1073/pnas.1200913109
7. Kazazian, H. (2004). Mobile elements: Drivers of genome evolution. *Science, 303*(5664), 1626-1632. doi:10.1126/science.1089670
8. Stocking, C., & Kozak, C. A. (2008). Murine endogenous retroviruses. *Cellular and Molecular Life Sciences, 65*(21), 3383-3398. doi:10.1007/s00018-008-8497-0
9. Rebollo, R., Farivar, S., & Mager, D. L. (2012). C-GATE - catalogue of genes affected by transposable elements. *Mobile Dna, 3*, 9. doi:10.1186/1759-8753-3-9
10. Maksakova, I. A., Romanish, M. T., Gagnier, L., Dunn, C. A., de Lagemaat, L. N. v., & Mager, D. L. (2006). Retroviral elements and their hosts: Insertional mutagenesis in the mouse germ line. *Plos Genetics, 2*(1), 1-10. doi:10.1371/journal.pgen.0020002
11. Vasicek, T. J., Zeng, L., Guan, X. J., Zhang, T., Costantini, F., & Tilghman, S. M. (1997). Two dominant mutations in the mouse fused gene are the result of transposon insertions. *Genetics, 147*(2), 777-786
12. Blake JA, Eppig JT, Kadin JA, Richardson JE, Smith CL, Bult CJ, and the Mouse Genome Database Group. 2017. Mouse Genome Database (MGD)-2017: community knowledge resource for the laboratory mouse. Nucl. Acids Res. 2017 Jan. 4;45 (D1): D723-D729.
13. Dickies, M. M. (1962). A new viable yellow mutation in house mouse. *Journal of Heredity, 53*(2), 84-&.
14. Jern, P., & Coffin, J. M. (2008). Effects of retroviruses on host genome function. *Annual Review of Genetics, 42*, 709-732. doi:10.1146/annurev.genet.42.110807.091501
15. Belshaw, R., Dawson, A. L. A., Woolven-Allen, J., Redding, J., Burt, A., & Tristem, M. (2005). Genomewide screening reveals high levels of insertional polymorphism in the human endogenous retrovirus family HERV-K(HML2): Implications for present-day activity. *Journal of Virology, 79*(19), 12507-12514. doi:10.1128/JVI.79.19.12507-12507-12514.2005

16. Yap, M. W., Colbeck, E., Ellis, S. A., & Stoye, J. P. (2014). Evolution of the retroviral restriction gene Fv1: Inhibition of non-MLV retroviruses. *Plos Pathogens, 10*(3), e1003968. doi:10.1371/journal.ppat.1003968

17. Gogvadze, E., & Buzdin, A. (2009). Retroelements and their impact on genome evolution and functioning. *Cellular and Molecular Life Sciences, 66*(23), 3727-3742. doi:10.1007/s00018-009-0107-2

18. Cohen, C. J., Lock, W. M., & Mager, D. L. (2009). Endogenous retroviral LTRs as promoters for human genes: A critical assessment. *Gene, 448*(2), 105-114. doi:10.1016/j.gene.2009.06.020

19. Thompson, P. J., Macfarlan, T. S., & Lorincz, M. C. (2016). Long terminal repeats: From parasitic elements to building blocks of the transcriptional regulatory repertoire. *Molecular Cell, 62*(5), 766-776. doi:10.1016/j.molcel.2016.03.029

20. van de Lagemaat, L. N., Medstrand, P., & Mager, D. L. (2006). Multiple effects govern endogenous retrovirus survival patterns in human gene introns. *Genome Biology, 7*(9), 86. doi:10.1186/gb-2006-7-9-r86

21. Hughes, D. C. (2001). Alternative splicing of the human VEGFGR-3/FLT4 gene as a consequence of an integrated human endogenous retrovirus. *Journal of Molecular Evolution, 53*(2), 77-79.

22. Mager, D. L., Hunter, D. G., Schertzer, M., & Freeman, J. D. (1999). Endogenous retroviruses provide the primary polyadenylation signal for two new human genes (HHLA2 and HHLA3). *Genomics, 59*(3), 255-263. doi:10.1006/geno.1999.5877

23. Mita, P., & Boeke, J. D. (2016). How retrotransposons shape genome regulation. *Current Opinion in Genetics & Development, 37*, 90-100. doi:http://dx.doi.org.ezproxy.library.ubc.ca/10.1016/j.gde.2016.01.001

24. Sharif, J., Shinkai, Y., & Koseki, H. (2013). Is there a role for endogenous retroviruses to mediate long-term adaptive phenotypic response upon environmental inputs? *Philosophical Transactions of the Royal Society B-Biological Sciences, 368*(1609), 20110340. doi:10.1098/rstb.2011.0340

25. Faulkner, G. J., Kimura, Y., Daub, C. O., Wani, S., Plessy, C., Irvine, K. M., . . . Carninci, P. (2009). The regulated retrotransposon transcriptome of mammalian cells. *Nature Genetics, 41*(5), 563-571. doi:10.1038/ng.368

26. van de Lagemaat, L., Landry, J., Mager, D., & Medstrand, P. (2003). Transposable elements in mammals promote regulatory variation and diversification of genes with specialized functions. *Trends in Genetics, 19*(10), 530-536. doi:10.1016/j.tig.2003.08.004

27. Rebollo, R., Miceli-Royer, K., Zhang, Y., Farivar, S., Gagnier, L., & Mager, D. L. (2012). Epigenetic interplay between mouse endogenous retroviruses and host genes. *Genome Biology, 13*(10), R89. doi:10.1186/gb-2012-13-10-R89

28. Grow, E. J., Flynn, R. A., Chavez, S. L., Bayless, N. L., Wossidlo, M., Wesche, D. J., . . . Wysocka, J. (2015). Intrinsic retroviral reactivation in human preimplantation embryos and pluripotent cells. *Nature, 522*(7555), 221-+. doi:10.1038/nature14308

29. Goeke, J., Lu, X., Chan, Y., Ng, H., Ly, L., Sachs, F., & Szczerbinska, I. (2015). Dynamic transcription of distinct classes of endogenous retroviral elements marks specific populations of early human embryonic cells. *Cell Stem Cell, 16*(2), 135-141. doi:10.1016/j.stem.2015.01.005

30. Peaston, A., Evsikov, A., Graber, J., de Vries, W., Holbrook, A., Solter, D., & Knowles, B. (2004). Retrotransposons regulate host genes in mouse oocytes and preimplantation embryos. *Developmental Cell, 7*(4), 597-606. doi:10.1016/j.devcel.2004.09.004

31. Stavenhagen, J. B., & Robins, D. M. (1988). An ancient provirus has imposed androgen regulation on the adjacent mouse sex-limited protein gene. *Cell, 55*(2), 247-254. doi:10.1016/0092-8674(88)90047-5

32. Meisler, M. H., & Ting, C. N. (1993). The remarkable evolutionary history of the human amylase genes. *Critical Reviews in Oral Biology & Medicine, 4*(3-4), 503-509.

33. Sundaram, V., Cheng, Y., Ma, Z., Li, D., Xing, X., Edge, P., . . . Wang, T. (2014). Widespread contribution of transposable elements to the innovation of gene regulatory networks. *Genome Research, 24*(12), 1963-1976. doi:10.1101/gr.168872.113

34. Long, Q., Bengra, C., Li, C., Kutlar, F., & Tuan, D. (1998). A long terminal repeat of the human endogenous retrovirus ERV-9 is located in the 5′ boundary area of the human β-globin locus control region. *Genomics, 54*(3), 542-555. doi:http://dx.doi.org.ezproxy.library.ubc.ca/10.1006/geno.1998.5608

35. Cao, A., & Moi, P. (2002). Regulation of the globin genes. *Pediatric Research, 51*(4), 415-421.

36. Haig, D. (2012). Retroviruses and the placenta. *Current Biology, 22*(15), R609-R613. doi:10.1016/j.cub.2012.06.002

37. Dupressoir, A., Lavialle, C., & Heidmann, T. (2012). From ancestral infectious retroviruses to bona fide cellular genes: Role of the captured syncytins in placentation. *Placenta, 33*(9), 663-671. doi:10.1016/j.placenta.2012.05.005

38. Cornelis, G., Vernochet, C., Carradec, Q., Souquere, S., Mulot, B., Catzeflis, F., . . . Heidmann, T. (2015). Retroviral envelope gene captures and syncytin exaptation for placentation in marsupials. *Proceedings of the National Academy of Sciences of the United States of America, 112*(5), E487-E496. doi:10.1073/pnas.1417000112

39. Romanish, M. T., Lock, W. M., van de Lagemaat, L. N., Dunn, C. A., & Mager, D. L. (2007). Repeated recruitment of LTR retrotransposons as promoters by the anti-apoptotic locus NAIP during mammalian evolution. *Plos Genetics, 3*(1), e10. doi:10.1371/journal.pgen.0030010

40. Tsuritani, K., Irie, T., Yamashita, R., Sakakibara, Y., Wakaguri, H., Kanai, A., . . . Suzuki, Y. (2007). Distinct class of putative "non-conserved" promoters in humans: Comparative studies of alternative promoters of human and mouse genes. *Genome Research, 17*(7), 1005-1014. doi:10.1101/gr.6030107

41. Polavarapu, N., Marino-Ramirez, L., Landsman, D., McDonald, J., & Jordan, I. K. (2008). Evolutionary rates and patterns for human transcription factor binding sites derived from repetitive DNA. *BioMed Central Genomics, 9*, 226.

42. Kunarso, G., Chia, N., Jeyakani, J., Hwang, C., Lu, X., Chan, Y., . . . Bourque, G. (2010). Transposable elements have rewired the core regulatory network of human embryonic stem cells. *Nature Genetics, 42*(7), 631-U111. doi:10.1038/ng.600

43. Yoder, J. A., Walsh, C. P., & Bestor, T. H. (1997). Cytosine methylation and the ecology of intragenomic parasites. *Trends in Genetics, 13*(8), 335-340. doi:10.1016/S0168-9525(97)01181-5

44. Leung, D. C., & Lorincz, M. C. (2012). Silencing of endogenous retroviruses: When and why do histone marks predominate? *Trends in Biochemical Sciences, 37*(4), 127-133. doi:10.1016/j.tibs.2011.11.006

45. Maksakova, I. A., Mager, D. L., & Reiss, D. (2008). Keeping active endogenous retroviral-like elements in check: The epigenetic perspective. *Cellular and Molecular Life Sciences, 65*(21), 3329--3347. doi:10.1007/s00018-008-8494-3

46. Rowe, H. M., & Trono, D. (2011). Dynamic control of endogenous retroviruses during development. *Virology, 411*(2), 273-287. doi:10.1016/j.virol.2010.12.007

47. Friedli, M., & Trono, D. (2015). The developmental control of transposable elements and the evolution of higher species. *Annual Review of Cell and Developmental Biology, Vol 31, 31*, 429-451. doi:10.1146/annurev-cellbio-100814-125514

48. Ecco, G., Cassano, M., Kauzlaric, A., Duc, J., Coluccio, A., Offner, S., . . . Trono, D. (2016). Transposable elements and their KRAB-ZFP controllers regulate gene expression in adult tissues. *Developmental Cell, 36*(6), 611-623. doi:10.1016/j.devcel.2016.02.024

49. Rowe, H. M., Kapopoulou, A., Corsinotti, A., Fasching, L., Macfarlan, T. S., Tarabay, Y., . . . Trono, D. (2013). TRIM28 repression of retrotransposon-based enhancers is necessary to preserve transcriptional dynamics in embryonic stem cells. *Genome Research, 23*(3), 452-461. doi:10.1101/gr.147678.112

50. Thomas, J. H., & Schneider, S. (2011). Coevolution of retroelements and tandem zinc finger genes. *Genome Research, 21*(11), 1800-1812. doi:10.1101/gr.121749.111

51. Wolf, G., Greenberg, D., & Macfarlan, T. S. (2015). Spotting the enemy within: Targeted silencing of foreign DNA in mammalian genomes by the kruppel-associated box zinc finger protein family. *Mobile Dna, 6*, 17. doi:10.1186/s13100-015-0050-8

52. Jakobsson, J., Cordero, M. I., Bisaz, R., Groner, A. C., Busskamp, V., Bensadoun, J., . . . Trono, D. (2008). KAP1-mediated epigenetic repression in the forebrain modulates behavioral vulnerability to stress. *Neuron, 60*(5), 818-831. doi:10.1016/j.neuron.2008.09.036

53. Rakyan, V. K., Blewitt, M. E., Druker, R., Preis, J. I., & Whitelaw, E. (2002). Metastable epialleles in mammals. *Trends in Genetics, 18*(7), 348-351. doi:10.1016/S0168-9525(02)02709-9

54. Dolinoy, D. C. (2008). The agouti mouse model: An epigenetic biosensor for nutritional and environmental alterations on the fetal epigenome. *Nutrition Reviews, 66*(8), S7-S11. doi:10.1111/j.1753-4887.2008.00056.x

55. Chuong, E. B., Elde, N. C., & Feschotte, C. (2016). Regulatory evolution of innate immunity through co-option of endogenous retroviruses. *Science, 351*(6277), 1083-1087. doi:10.1126/science.aad5497

56. Lynch, V. J. (2016). A copy-and-paste gene regulatory network. *Science, 351*(6277), 1029-1030. doi:10.1126/science.aaf2977

57. Nellaker, C., Keane, T. M., Yalcin, B., Wong, K., Agam, A., Belgard, T. G., . . . Ponting, C. P. (2012). The genomic landscape shaped by selection on transposable elements across 18 mouse strains. *Genome Biology, 13*(6), R45. doi:10.1186/gb-2012-13-6-r45

58. Zhang, Y., Maksakova, I. A., Gagnier, L., de Lagemaat, L. N. v., & Mager, D. L. (2008). Genome-wide assessments reveal extremely high levels of polymorphism of two active families of mouse endogenous retroviral elements. *Plos Genetics, 4*(2), e1000007. doi:10.1371/journal.pgen.1000007

59. Qin, C., Wang, Z., Shang, J., Bekkari, K., Liu, R., Pacchione, S., . . . Storer, R. D. (2010). Intracisternal A particle genes: Distribution in the mouse genome, active subtypes, and potential roles as species-specific mediators of susceptibility to cancer. *Molecular Carcinogenesis, 49*(1), 54-67. doi:10.1002/mc.20576

60. Saito, E., Keng, V. W., Takeda, J., & Horie, K. (2008). Translation from nonautonomous type IAP retrotransposon is a critical determinant of transposition activity: Implication for retrotransposon-mediated genome evolution. *Genome Research, 18*(6), 859-868. doi:10.1101/gr.069310.107

61. Christy, R. J., Brown, A. R., Gourlie, B. B., & Huang, R. C. C. (1985). Nucleotide-sequences of murine intracisternal A-particle gene ltrs have extensive variability within the R-region. *Nucleic Acids Research, 13*(1), 289-302. doi:10.1093/nar/13.1.289

62. Kuff, E., & Leuders, K. (1988). The intracisternal A-particle gene family: Structure and functional aspects. *Advances in Cancer Research, 51*, 183-276.

63. TEMIN, H. (1981). Structure, variation and synthesis of retrovirus long terminal repeat. *Cell, 27*(1), 1-3. doi:10.1016/0092-8674(81)90353-6

64. TEMIN, H. (1982). Function of the retrovirus long terminal repeat. *Cell, 28*(1), 3-5. doi:10.1016/0092-8674(82)90367-1

65. Christy, R. J., & Huang, R. C. C. (1988). Functional-analysis of the long terminal repeats of intracisternal A-particle genes - sequences within the U3-region determine both the efficiency and direction of promoter activity. *Molecular and Cellular Biology, 8*(3), 1093-1102.

66. Zierler, M., Christy, R. J., & Huang, R. C. C. (1992). Nuclear-protein binding to the 5' enhancer region of the intracisternal-a particle long terminal repeat. *Journal of Biological Chemistry, 267*(29), 21200-21206.

67. Falzon, M., & Kuff, E. L. (1988). Multiple protein-binding sites in an intracisternal A-particle long terminal repeat. *Journal of Virology, 62*(11), 4070-4077.

68. Falzon, M., & Kuff, E. L. (1989). Isolation and characterization of a protein-fraction that binds to enhancer core sequences in intracisternal A-particle long terminal repeats. *Journal of Biological Chemistry, 264*(36), 21915-21922.

69. Falzon, M., & Kuff, E. L. (1991). Binding of the transcription factor ebp-80 mediates the methylation response of an intracisternal A-particle long terminal repeat promoter. *Molecular and Cellular Biology, 11*(1), 117-125.

70. Fasching, L., Kapopoulou, A., Sachdeva, R., Petri, R., Jonsson, M. E., Manne, C., . . . Jakobsson, J. (2015). TRIM28 represses transcription of endogenous retroviruses in neural progenitor cells. *Cell Reports, 10*(1), 20-28. doi:10.1016/j.celrep.2014.12.004

71. Kapitonov, V. V., & Jurka, J. (2008). A universal classification of eukaryotic transposable elements implemented in repbase. *Nature Reviews Genetics, 9*(5) doi:10.1038/nrg2165-c1

72. Trinklein, N., Aldred, S., Hartman, S., Schroeder, D., Otillar, R., & Myers, R. (2004). An abundance of bidirectional promoters in the human genome. *Genome Research, 14*, 62-66.

73. Danino, Y. M., Even, D., Ideses, D., & Juven-Gershon, T. (2015). The core promoter: At the heart of gene expression. *Biochimica Et Biophysica Acta-Gene Regulatory Mechanisms, 1849*(8), 1116-1131. doi:10.1016/j.bbagrm.2015.04.003

74. Engstrom, P. G., Suzuki, H., Ninomiya, N., Akalin, A., Sessa, L., Lavorgna, G., . . . Lipovich, L. (2006). Complex loci in human and mouse genomes. *Plos Genetics, 2*(4), 564-577. doi:10.1371/journal.pgen.0020047

75. Orekhova, A. S., & Rubtsov, P. M. (2013). Bidirectional promoters in the transcription of mammalian genomes. *Biochemistry-Moscow, 78*(4), 335-341. doi:10.1134/S0006297913040020

76. Lin, S., Zhang, L., Luo, W., & Zhang, X. (2016). Characteristics of antisense transcript promoters and the regulation of their activity. *International Journal of Molecular Sciences, 17*(9)

77. Yang, M., Koehly, L., & Elnitski, L. (2007). Comprehensive annotation of bidirectional promoters identifies co-regulation among breast and ovarian cancer genes. *PLoS Computation Biology, 3*(4), e72. doi:10.1371/journal.pcbi.0030072.eor

78. Polson, A., Durrett, E., & Reisman, D. (2011). A bidirectional promoter reporter vector for the analysis of the p53/WDR79 dual regulatory element. *Plasmid, 66*(3), 169-179. doi:10.1016/j.plasmid.2011.08.004

79. Farlex. (2017). The free dictionary. Retrieved on January 20, 201 from: http://www.thefreedictionary.com/

80. Promega Corporation. (2017). Promega. Retrieved on January 22, 2017 from: https://www.promega.ca/resources/protocols/technical-manuals/0/dual-glo-luciferase-assay-system-protocol/

81. Wu, J. Q., & Snyder, M. (2008). RNA polymerase II stalling: Loading at the start prepares genes for a sprint. *Genome Biology, 9*(5), 220. doi:10.1186/gb-2008-9-5-220

82. Wei, W., Pelechano, V., Jaervelin, A. I., & Steinmetz, L. M. (2011). Functional consequences of bidirectional promoters. *Trends in Genetics, 27*(7), 267-276. doi:10.1016/j.tig.2011.04.002

83. Natoli, G., & Andrau, J. (2012). Noncoding transcription at enhancers: General principles and functional models. *Annual Review of Genetics, Vol 46, 46*, 1-19. doi:10.1146/annurev-genet-110711-155459

84. Thermo Fisher Scientific Incorporated. (2017). ThermoFisher scientific. Retrieved on January 22, 2017 from: https://www.thermofisher.com/order/catalog/product/T1012

85. Dolinoy, D. C. (2008). The agouti mouse model: An epigenetic biosensor for nutritional and environmental alterations on the fetal epigenome. *Nutrition Reviews, 66*(8), S7-S11. doi:10.1111/j.1753-4887.2008.00056.x

86. Michaud, E. J., Vanvugt, M. J., Bultman, S. J., Sweet, H. O., Davisson, M. T., & Woychik, R. P. (1994). Differential expression of a new dominant agouti allele (A(iapy)) is correlated with methylation state and is influenced by parental lineage. *Genes & Development, 8*(12), 1463-1472. doi:10.1101/gad.8.12.1463

87. PERRY, W., COPELAND, N., & JENKINS, N. (1994). The molecular-basis for dominant yellow agouti coat color mutations. *Bioessays, 16*(10), 705-707. doi:10.1002/bies.950161002

88. Juriloff, D. M., Harris, M. J., Mager, D. L., & Gagnier, L. (2014). Epigenetic mechanism causes Wnt9b deficiency and nonsyndromic cleft lip and palate in the A/WySn mouse strain. *Birth Defects Research Part A-Clinical and Molecular Teratology, 100*(10), 772-788. doi:10.1002/bdra.23320

89. Kuster, J., Guarnieri, M., Ault, J., Flaherty, L., & Swiatek, P. (1997). IAP insertion in the murine LamB3 gene results in junctional epidermolysis bullosa. *Mammalian Genome, 8*(9), 673-681. doi:10.1007/s003359900535

90. Druker, R., Bruxner, T., Lehrbach, N., & Whitelaw, E. (2004). Complex patterns of transcription at the insertion site of a retrotransposon in the mouse. *Nucleic Acids Research, 32*(19), 5800-5808. doi:10.1093/nar/gkh914

91. Ekram, M. B., Kang, K., Kim, H., & Kim, J. (2012). Retrotransposons as a major source of epigenetic variations in the mammalian genome. *Epigenetics, 7*(4), 370-382.

92. Aberdam, D., Negreanu, V., Sachs, L., & Blatt, C. (1991). The oncogenic potential of an activated hox-2.4 homeobox gene in mouse fibroblasts. *Molecular and Cellular Biology, 11*(1), 554-557.

93. Bendavid, L., Aberdam, D., Sachs, L., & Blatt, C. (1991). A deletion and a rearrangement distinguish between the intracisternal A-particle of hox-2.4 and that of interleukin-3 in the same leukemic-cells. *Virology, 182*(1), 382-387. doi:10.1016/0042-6822(91)90686-6

94. Kongsuwan, K., Allen, J., & Adams, J. M. (1989). Expression of hox-2.4 homeobox gene directed by proviral insertion in a myeloid-leukemia. *Nucleic Acids Research, 17*(5), 1881-1892. doi:10.1093/nar/17.5.1881

95. Ymer, S., Tucker, W. Q. J., Campbell, H. D., & Young, I. G. (1986). Nucleotide-sequence of the intracisternal A-particle genome inserted 5' to the interleukin-3 gene of the leukemia-cell line wehi-3b. *Nucleic Acids Research, 14*(14), 5901-5918. doi:10.1093/nar/14.14.5901

96. Leupin, O., Attanasio, C., Marguerat, S., Tapernoux, M., Antonarakis, S. E., & Conrad, B. (2005). Transcriptional activation by bidirectional RNA polymerase II elongation over a silent promoter. *EMBO Reports, 6*(10), 956-960. doi:10.1038/sj.embor.7400502

97. DUHL, D., VRIELING, H., MILLER, K., WOLFF, G., & BARSH, G. (1994). Neomorphic agouti mutations in obese yellow mice. *Nature Genetics, 8*(1), 59-65. doi:10.1038/ng0994-59

98. Juven-Gershon, T., & Kadonaga, J. T. (2010). Regulation of gene expression via the core promoter and the basal transcriptional machinery. *Developmental Biology, 339*(2), 225-229. doi:10.1016/j.ydbio.2009.08.009

99. Xavier Messeguer, Ruth Escudero, Domènec Farré, Oscar Nuñez, Javier Martínez, M.Mar Albà. (2002) PROMO: detection of known transcription regulatory elements using species-tailored searches. Bioinformatics, 18, 2, 333-334.

100. Kwon AT, Arenillas DJ, Worsley Hunt R, Wasserman WW. (2012) oPOSSUM-3: advanced analysis of regulatory motif over-representation across genes or ChIP-Seq datasets. G3 epub.2(9):987-1002. doi: 10.1534/g3.112.003202.

101. Rebhan, M., Chalifa-Caspi, V., Prilusky, J., & Lancet, D. (1997). GeneCards: Integrating information about genes, proteins and diseases. Trends in Genetics, 13(163)

102. Aguilera, C., Nakagawa, K., Sancho, R., Chakraborty, A., Hendrich, B., & Behrens, A. (2011). c-jun N-terminal phosphorylation antagonises recruitment of the Mbd3/NuRD repressor complex. *Nature, 469*(7329), 231-235. doi:10.1038/nature09607

103. Guo, W., Keckesova, Z., Donaher, J. L., Shibue, T., Tischler, V., Reinhardt, F., . . . Weinberg, R. A. (2012). Slug and Sox9 cooperatively determine the mammary stem cell state. *Cell, 148*(5), 1015-1028. doi:10.1016/j.cell.2012.02.008

104. Hu, C., Sataur, A., Wang, L., Chen, H., & Simon, M. C. (2007). The N-terminal transactivation domain confers target gene specificity of hypoxia-inducible factors HIF-1 alpha and HIF-2 alpha. *Molecular Biology of the Cell, 18*(11), 4528-4542. doi:10.1091/mbc.E06-05-0419

105. Roczniak-Ferguson, A., Petit, C. S., Froehlich, F., Qian, S., Ky, J., Angarola, B., . . . Ferguson, S. M. (2012). The transcription factor TFEB links mTORC1 signaling to transcriptional control of lysosome homeostasis. *Science Signaling, 5*(228), ra42. doi:10.1126/scisignal.2002790

106. Hojman-Montes de Oca, F., Dianoux, L., Peries, J., & Emanoil-Ravicovitch, R. (1983). Intracisternal A particles: RNA expression and DNA methylation in murine teratocarcinoma cell lines. *Journal of Virology, 46*(1), 307-10.

107. Barbeau, B., & Mesnard, J. (2015). Does chronic infection in retroviruses have a sense? *Trends in Microbiology, 23*(6), 367-375. doi:10.1016/j.tim.2015.01.009

108. CANAANI, E., DREAZEN, O., KLAR, A., RECHAVI, G., RAM, D., COHEN, J., & GIVOL, D. (1983). Activation of the C-mos oncogene in a mouse plasmacytoma by insertion of an endogenous intracisternal A-particle genome. *Proceedings of the National Academy of Sciences of the United States of America-Biological Sciences, 80*(23), 7118-7122. doi:10.1073/pnas.80.23.7118

109. Lee, J., Haruna, T., Ishimoto, A., Honjo, T., & Yanagawa, S. (1999). Intracisternal type A particle-mediated activation of the Notch4/int3 gene in a mouse mammary tumor: Generation of truncated Notch4/int3 mRNAs by retroviral splicing events. *Journal of Virology, 73*(6), 5166-5171.

110. BLANKENSTEIN, T., QIN, Z., LI, W., & DIAMANTSTEIN, T. (1990). Dna rearrangement and constitutive expression of the interleukin 6-gene in a mouse plasmacytoma. *Journal of Experimental Medicine, 171*(3), 965-970. doi:10.1084/jem.171.3.965

111. DUHRSEN, U., STAHL, J., & GOUGH, N. (1990). Invivo transformation of factor-dependent hematopoietic-cells - role of intracisternal A-particle transposition for growth-factor gene activation. *Embo Journal, 9*(4), 1087-1096.

112. Ishihara, H., Tanaka, I., Wan, H., Nojima, K., & Yoshida, K. (2004). Retrotransposition of limited deletion type of intracisternal A-particle elements in the myeloid leukemia clls of C3H/He mice. *Journal of Radiation Research, 45*(1), 25-32. doi:10.1269/jrr.45.25

113. BRIGLE, K., WESTIN, E., HOUGHTON, M., & GOLDMAN, I. (1992). Insertion of an intracisternal-a particle within the 5'-regulatory region of a gene encoding folate-binding protein in L1210 leukemia-cells in response to low folate selection - association with increased protein expression. *Journal of Biological Chemistry, 267*(31), 22351-22355.

114. TANAKA, I., & ISHIHARA, H. (1995). Unusual long target duplication by insertion of intracisternal A-particle element in radiation-induced acute myeloid-leukemia cells in mouse. *FEBS Letters, 376*(3), 146-150. doi:10.1016/0014-5793(95)01262-2

115. Gardner, J., Wildenberg, S., Keiper, N., Novak, E., Rusiniak, M., Swank, R., . . . Brilliant, M. (1997). The mouse pale ear (ep) mutation is the homologue of human hermansky-pudlak syndrome. *Proceedings of the National Academy of Sciences of the United States of America, 94*(17), 9238-9243. doi:10.1073/pnas.94.17.9238

116. Kantheti, P., Qiao, X., Diaz, M., Peden, A., Meyer, G., Carskadon, S., . . . Burmeister, M. (1998). Mutation in AP-3 delta in the mocha mouse links endosomal transport to storage deficiency in platelets, melanosomes, and synaptic vesicles. *Neuron, 21*(1), 111-122. doi:10.1016/S0896-6273(00)80519-X

117. The Jackson Laboratory. (2017). The jackson laboratory. Retrieved on January 25, 2017 from: https://www.jax.org/

118. Blake, J., Eppig J., Kadin, J., Richardson, J., Smith, C., Bult, C., and the Mouse Genome Database Group. (2017). Mouse Genome Database (MGD)-2017: community knowledge resource for the laboratory mouse. Nucl. Acids Res. 4;45 (D1): D723-D729.

119. Finger, J., Smith, C., Hayamizu, T., McCright, I., Xu, J., Law, M., Shaw, D., Baldarelli, R., Beal, J., Blodgett, O., Campbell, J., Corbani, L., Lewis, J., Forthofer, K., Frost, P., Giannatto, S., Hutchins, L., Miers, D., Motenko, H., Stone, K., Eppig, J., Kadin, J., Richardson, J., Ringwald, M. (2017). The mouse Gene Expression Database (GXD): 2017 update. Nucleic Acids Res. 4;45 (D1): D730-D736.

120. Aguilera C, Nakagawa K, Sancho R, Chakraborty A, Hendrich B, Behrens A. (2011.) c-Jun N-terminal phosphorylation antagonises recruitment of the Mbd3/NuRD repressor complex. Nature. 469(7329):231-5. doi: 10.1038/nature09607. Epub 2011 Jan 2. 10.1038/nature09607 PubMed 21196933

121. Roczniak-Ferguson A, Petit CS, Froehlich F, Qian S, Ky J, Angarola B, Walther TC, Ferguson SM. (2012.) The Transcription Factor TFEB Links mTORC1 Signaling to Transcriptional Control of Lysosome Homeostasis. Sci Signal. 5(228):ra42. 10.1126/scisignal.2002790 PubMed 22692423

122. Guo W, Keckesova Z, Donaher JL, Shibue T, Tischler V, Reinhardt F, Itzkovitz S, Noske A, Zurrer-Hardi U, Bell G, Tam WL, Mani SA, van Oudenaarden A, Weinberg RA. (2012.)Slug and Sox9 cooperatively determine the mammary stem cell state. Cell. 148(5):1015-28. 10.1016/j.cell.2012.02.008 PubMed 22385965

123. Hu CJ, Sataur A, Wang L, Chen H, Simon MC. (2007.) The N-terminal transactivation domain confers target gene specificity of hypoxia-inducible factors HIF-1alpha and HIF-2alpha. Mol Biol Cell. 18(11):4528-42. Epub 2007 Sep 5. 10.1091/mbc.E06-05-0419 PubMed 17804822

124. ThermoFIsher Scientific. (2017). The history of PCR. Retrieved on February 18, 2017 from: https://www.thermofisher.com/ca/en/home/brands/thermo-scientific/molecular-biology/molecular-biology-learning-center/molecular-biology-resource-library/spotlight-articles/history-pcr.html

125. Promega Corporation. (2017). Promega. Retrieved on February 20, 2017 from: https://www.promega.com/-/media/files/resources/protocols/technical-manuals/0/pgl3-luciferase-reporter-vectors-protocol.pdf

126. Hampton, R. E., Havel, J. E. (2006) Introductory Biological Statistics. Second Edition, Waveland Press Inc. USA. p157.

# Appendices

## Appendix A   Primers and Fragments

| Name | Primer | Used for |
|---|---|---|
| Wnt9bLTR-F | CTGAAACCCGGGGATTGTTATTCGACGC | Frags |
| Wnt9bLTR-R | TGTCTACCCGGGGATTGTTATTCGACGC | Frags |
| LamB3LTR_L | CTGAAACCCGGGTGTTGGGAGCCGCC | Frags |
| LamB3LTR_D1 | CTTCAACCATGGCACTTAGAACACAGGATGTCAGC | Frags |
| LamB3LTR_D2 | CTGAAACCATGGGGGGAAGGCAGAGCACAAGTA | Frags |
| LamB3LTR_D3 | CTGAAACCATGGGAGCAAGAAGCAAGAGAGAGA | Frags |
| LamB3LTR_D4 | CTGAAACCCGGGCTACTTGTGCTCTGCCTTCC | Frags |
| LamB3LTR_D5 | TGTCTACCCGGGCGGGGTTTTCGTTTTCTCTC | Frags |
| LamB3LTR_D6 | CTGAAACCCGGGCCGCAGAAGATTCTGGTCTG | Frags |
| LamB3LTR_R | CTTGGTCCATGGTGTTATTCGACGCGTTCTCA | Frags |
| LamB3LTR_D7 | ATTTCACCATGGGGACGTGTCACTCCCTGA | Frags |
| LamB3LTR_D8 | ATTGATCCATGGGGACGTGTCACTCCCTGA | Frags |
| LamB3LTR_D10 | TGTTTTCCCGGGCTTAAGAGGGACGGGGTTTT | Frags |
| pLucRLuc_SeqL | GCCTTATGCAGTTGCTCTCC | Seq. |
| pLucRLuc_SeqR | TGATACTTACCTGCCCAGTGC | Seq. |
| del2096-antisense_2a | GAGCACAAGTAGTCATAAATACCCTTGGCTCATGCG | P.M.s |
| de12096_2a | CGCATGAGCCAAGGGTATTTATGACTACTTGTGCTC | P.M.s |
| t2100c_3 | CAGAGCACAAGTAGTCGTAAGATACCCTTGGCTC | P.M.s |
| t2100c_4 | GAGCCAAGGGTATCTTACGACTACTTGTGCTCTG | P.M.s |
| del2096-antisense_2b | CACAAGTAGTCGTAAATACCCTTGGCTCATGCGC | P.M.s |
| del2096_2b | GCGCATGAGCCAAGGGTATTTACGACTACTTGTG | P.M.s |
| del2096-antisense_3 | CACAAGTAGTCGTAAATACCCTTGGCACATGCGC | P.M.s |
| del2096_3 | GCGCATGTGCCAAGGGTATTTACGACTACTTGTG | P.M.s |

P.M.s=Point Mutations; Frags=Fragment Generation; Seq=Sequencing; LamB3=IL3 LTR

| Fragment | Primer 1 | Primer 2 |
|---|---|---|
| A | LamB3LTR_L | LamB3LTR_D1 |
| B | LamB3LTR_L | LamB3LTR_D2 |
| C | LamB3LTR_L | LamB3LTR_D3 |
| D | LamB3LTR_L | LamB3LTR_R |
| E | LamB3LTR_D4 | LamB3LTR_R |
| F | LamB3LTR_D5 | LamB3LTR_R |
| G | LamB3LTR_D6 | LamB3LTR_R |
| H | LamB3LTR_D4 | LamB3LTR_D7 |
| I | LamB3LTR_D4 | LamB3LTR_D8 |
| J | LamB3LTR_D10 | LamB3LTR_D3 |
| K | LamB3LTR_D4 | LamB3LTR_D3 |
| Wnt9bLTR | Wnt9bLTR-F | Wnt9bLTR-R |

**Appendix B  Oligonucleotides and Sequences of Interest**

Sequence of Interest 1 (SOI1)

Oligo 1:        CATGGTCATAAGATACCCTTGGCACATGCGCAGATTATTTGTTTACC

Oligo 2:        CCGGGGTAAACAAATAATCTGCGCATGTGCCAAGGGTATCTTATGAC

Sequence of Interest 2 (SOI2)

Oligo 1:        CATGGGGACGTGTCACTCCCTGATTGGCTGCAGCCCC

Oligo 2:        CCGGGGGGCTGCAGCCAATCAGGGAGTGACACGTCC

Sequence of Interest 3 (SOI3)

Oligo 1:        CATGGGAGCATTTCCTTCGCCTCC

Oligo 2:        CCGGGGAGGCGAAGGAAATGCTCC