Characterization of the mechanisms of HERV-K (HML-2) transcription during human mammary epithelial cell transformation

A thesis submitted by

Meagan Montesion

in partial fulfillment of the requirements for the degree of

PhD

in

Genetics

Tufts University

May 2017

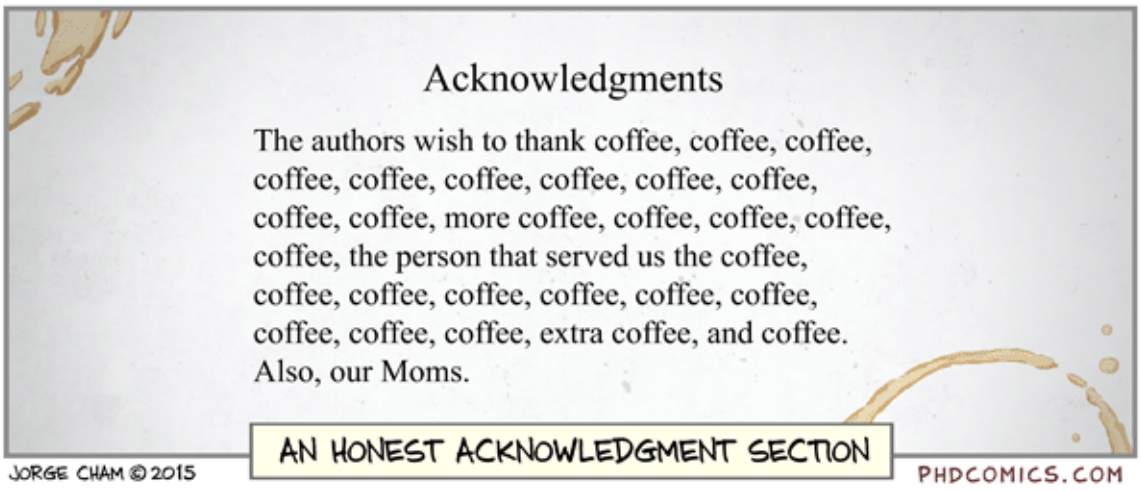
Advisor: John M. Coffin, PhD

Abstract

Increasing evidence suggests that repetitive elements may play a role in host gene regulation, particularly through the donation of alternative promoters, enhancers, splice sites, and termination signals. Elevated transcript expression of the endogenous retrovirus group HERV-K (HML-2) is seen in many human cancers, although the identity of the individual proviral loci contributing to this expression as well as their mechanism(s) of activation is unclear. Using a combination of reporter construct assays and RNA sequencing results, we characterized the HML-2 transcriptome and means of transcription in an *in vitro* model of human mammary epithelial cell transformation. RNA-Seq analysis showed transcription occurring through four different modes, with the majority of expression being in antisense orientation and from proviruses integrated within introns. Although we found two instances of LTR-driven provirus transcription, there was no evidence to suggest that these active 5' LTRs were directly influencing nearby host gene expression in the cell lines tested. Importantly, LTR-driven transcription was restricted to tumorigenic cells, suggesting that LTR promoter activity is dependent upon the transcriptional environment of a malignant cell. Using a transcription factor binding site prediction algorithm, we identified two unique binding sites on each 5' LTR of two highly active proviruses (3q12.3 and 11p15.4) that appeared to be associated with inducing promoter activity during neoplasia. Genomic analyses of the homologous proviruses in several non-human primate species indicated genetic drift in two of the binding sites, away from the ancestral sequence and towards the active form. Based on the sequences of 2,504 individuals from the 1000 Genomes Project, one of these sites was polymorphic in the human population with an allele frequency of 51%. These data

suggest that cell-specific transcription factors at least partly contribute to LTR promoter activity during transformation and that transcription factor binding site polymorphisms may be responsible for the differential HML-2 activity that is often seen amongst individuals and various disease states. These findings may provide implications for future studies investigating HML-2 as a target for immunotherapy or as a molecular biomarker of disease.

Acknowledgments



I must first thank John Coffin, my mentor throughout my time at Tufts, who was instrumental in getting me to this point and providing me the freedom and creativity to develop my own scientific skill set. To the remainder of my lab, both past and present – Neeru Bhardwaj, Mike Freeman, Joe Holloway, Farrah Roy, Zach Williams, and John Yoon – this experience would not have been the same without all of you. I will truly miss our troubleshooting powwows, cranberry harvests, labsgivings, and (at times) hours long coffee breaks. Additionally, special thank you to the barista at the local coffee shop who, towards the end of writing this dissertation, began asking me if I actually lived in the building upstairs.

Furthermore, I owe endless gratitude to my family. Their love and support has never wavered and I would not be the person I am today without their constant encouragement. I would also like to extend special thanks to Uri Bulow, Leah Graham, Sneha Borikar, Catherine Flynn, Kristin Meegan, and Sara MacKenzie. I could fill an entire second dissertation with stories of each friendship over the years and their heartfelt assurance that one day I would be "**Ph**inishe**D**".

With that, I'll end with an email from my sister, received after my first co-author publication. May we all have people in our lives that encourage us with at least 26 exclamation points at a time.

From: Amy Montesion
Date: Wednesday, March 11, 2015 at 10:50 PM
To: Meagan Montesion
Subject: Re: Paper

Meagan!!!!!!!!!!!! You germ maven you! Why do you only send this stuff to Dad?!!?!?! I have no idea what I just read and I thought one my eyes was going to explode but I loved it!!!! I probably got a provirus on this airplane I'm on right now. Good thing I now know that could stay in my DNA.

Way to go!! *****Virtual High Five*****!!!
Amy

Table of Contents

List of Tables

Chapter 1

Chapter 2

Chapter 3

Chapter 4

Chapter 5

List of Figures

Chapter 1

Chapter 3

Chapter 4

Chapter 5

List of Copyrighted Materials

Rambaut A, Posada D, Crandall KA, Holmes EC. 2004. The causes and consequences of HIV evolution. Nat Rev Genet **5:**52-61.
Figure 1 was reprinted by permission for Figure 1-1.

Swanstrom R, Wills JW. 1997. Synthesis, Assembly, and Processing of Viral Proteins. *In* Coffin JM, Hughes SH, Varmus HE (ed), Retroviruses, Cold Spring Harbor.
Figure 1 was reprinted by permission for Figure 1-2A.

Jern P, Coffin JM. 2008. Effects of retroviruses on host genome function. Annu Rev Genet **42:**709-732.
Figure 1 was reprinted by permission for Figure 1-2B.
Table 1 was reprinted by permission for Table 1-3.

Telesnitsky A, Goff SP. 1997. Reverse Transcriptase and the Generation of Retroviral DNA. *In* Coffin JM, Hughes SH, Varmus HE (ed), Retroviruses, Cold Spring Harbor.
Figure 2 was reprinted by permission for Figure 1-3.

Rosenberg N. 2011. Overview of Retrovirology. *In* Dudley J (ed), Retroviruses and insights into cancer. Springer, NY.
Figure 1.5 was reprinted by permission for Figure 1-4.

Beemon K, Rosenberg N. 2012. Mechanisms of Oncogenesis by Avian and Murine Retroviruses. *In* Robertson ES (ed), Cancer Associated Viruses. Springer.
Table 27.1 was reprinted by permission for Table 1-2.

Bannert N, Kurth R. 2006. The evolutionary dynamics of human endogenous retroviral families. Annu Rev Genomics Hum Genet **7:**149-173.
Figure 4 was reprinted by permission for Figure 1-5.

Perot P, Bolze P, Mallet F. 2012. From Viruses to Genes: Syncytins. *In* Witzany G (ed), Viruses: Essential Agents of Life. Springer.
Figure 1B was reprinted by permission for Figure 1-6.

Conley AB, Jordan IK. 2012. Endogenous Retroviruses and the Epigenome. *In* Witzany G (ed), Viruses: Essential Agents of Life. Springer.
Figure 4 was reprinted by permission for Figure 1-7A.
Figure 5 was reprinted by permission for Figure 1-7B.

Reynier F, Verjat T, Turrel F, Imbert PE, Marotte H, Mougin B, Miossec P. 2009. Increase in human endogenous retrovirus HERV-K (HML-2) viral load in active rheumatoid arthritis. Scand J Immunol **70:**295-299.
Figure 1 was reprinted by permission for Figure 1-8.

Subramanian RP, Wildschutte JH, Russo C, Coffin JM. 2011. Identification, characterization, and comparative genomic distribution of the HERV-K (HML-2) group of human endogenous retroviruses. Retrovirology **8:**90.
Figure 4 was reprinted by permission for Figure 1-9.

Butler JE, Kadonaga JT. 2002. The RNA polymerase II core promoter: a key component in the regulation of gene expression. Genes Dev **16:**2583-2592.
Figure 1 was reprinted by permission for Figure 1-10.

Bhardwaj N, Montesion M, Roy F, Coffin JM. 2015. Differential expression of HERV-K (HML-2) proviruses in cells and virions of the teratocarcinoma cell line Tera-1. Viruses **7:**939-968.
Figure 2B was reprinted by permission for Figure 5-1A.
Figure 5B was reprinted by permission for Figure 5-3A.
Figure 5C was reprinted by permission for Figure 5-3B.

Permission letters for all copyrighted materials are provided as a supplementary file to this document. Articles published by Subramanian *et al.* and Bhardwaj *et al.* are open access and do not require formal letters of permission for figure reproduction.

List of Abbreviations

5-hmC	5-hydroxymethylcytosine
5-mC	5-methylcytosine
A	adenine
AAA	polyadenine tail
AIDS	acquired immune deficiency syndrome
AFR	African
ALV	avian leukosis virus
Amp	ampicillin
AMR	Ad Mixed American
AMY1C	amylase, alpha 1C
ANOVA	analysis of variance
APOBEC3G	apolipoprotein B mRNA editing enzyme catalytic subunit 3G
ATCC	American Type Culture Collection
ATF	activating transcription factor
BAM	binary alignment/map
BLAST	basic local alignment search tool
BLAT	BLAST-like alignment tool
BLV	bovine leukemia virus
bp	base pair
BRCA1/2	breast cancer genes 1 and 2
BRE	TFIIB recognition element
C	cytosine

c-*onc*	cellular oncogene
CA	capsid
CAR	chimeric antigen receptor
CCBL1	cysteine conjugate-beta lyase 1
CD4	cluster of differentiation 4
cDNA	complementary DNA
chr	chromosome
CON	consensus
CpG	cytosine-phosphate-guanine
DCIS	ductal carcinoma in situ
DHRS4L1	dehydrogenase/reductase 4 like 1
DMEM	Dulbecco's Modified Eagle Medium
DNA	deoxyribonucleic acid
DNase	deoxyribonuclease
DPE	downstream promoter element
dTTP	deoxythymidine triphosphate
dUTP	deoxyuridine triphosphate
EAS	East Asian
Env	envelope
ER	estrogen receptor
ERBB2	avian erythroblastosis oncogene B-2
ERV	endogenous retrovirus
ESR1/2	estrogen receptor 1 and 2

EUR	European
FBS	fetal bovine serum
FeLV	feline leukemia virus
FLASh	Fast Length Adjustment of Short Reads
FOXA	forkhead box protein A
FPKM	fragments per kilobase of transcript per million mapped reads
Fr-MLV	Friend murine leukemia virus
FV	Friend virus
Fv1/4	Friend virus susceptibility 1 and 4
FV-P	polycythemia-inducing strain of Friend virus
G	guanine
Gag	group-specific antigen
GaLV	Gibbon ape leukemia virus
GAPDH	glyceraldehyde-3-phosphate dehydrogenase
gp55	glycoprotein 55
GRCh37	Genome Reference Consortium, human reference genome build 37
H3K9Ac	histone 3 lysine 9 acetylation
H3K27Ac	histone 3 lyine 27 acetylation
H3K27Me3	histone 3 lysine 27 trimethylation
HDAC	histone deacetylase
HER2	human epidermal growth factor receptor 2
HERV	human endogenous retrovirus
HERV-K	human endogenous retrovirus group K

hESC	human embryonic stem cell
HFV	human foamy virus
hg19	human reference genome build 19 (UCSC Genome Browser)
HHEX	hematopoietically expressed homeobox
HIV	human immunodeficiency virus
HIVEP1	HIV-1 enhancer binding protein 1
HLF	human hepatic leukemia factor
HME	human mammary epithelial
HMEC	human mammary epithelial cells
HML-2	human MMTV-like, group 2
HMLE	HMECs transformed with SV40 large T antigen
HOX-PBX	pre-B cell leukemia homeobox
HPV	human papillomavirus
hTERT	human telomerase reverse transcriptase
HTLV	human T-lymphotropic virus
IDT	Integrated DNA Technologies
IGV	Integrative Genomics Viewer
IK3	IKAROS protein 3
IKZF5	IKAROS family zinc finger 5
IL-2	interleukin-2
IL-2R	interleukin-2 receptor
IN	integrase
indel	insertion or deletion

Inr	initiator element
IRF5	interferon regulatory factor 5
JS	Jensen-Shannon
JSRV	Jaagsiekte sheep retrovirus
kb	kilobase
KoRV	Koala retrovirus
LEPR	leptin receptor
LINC	long intergenic noncoding RNA
log	logarithm
lncRNA	long non-coding RNA
LTR	long terminal repeat
Luc	firefly luciferase
MA	matrix
mAbs	monoclonal antibodies
MEGA	Molecular Evolutionary Genetics Analysis
MEGM	mammary epithelial cell growth medium
MEM	minimum essential medium
MERV	murine endogenous retrovirus
MHC	major histocompatibility complex
miRNA	microRNA
MLV	murine leukemia virus
MMTV	mouse mammary tumor virus
mRNA	messenger RNA

MUSCLE	multiple sequence comparison by log-expectation
mya	million years ago
myr	million years ago
NC	nucleocapsid
NCBI	National Center for Biotechnology Information
Neo	neomycin
neu	neuro/glioblastoma derived oncogene
NGS	next-generation sequencing
nm	nanometer
NONO	non-POU domain-containing, octamer-binding protein
NRL	neural retina leucine zipper
n.s.	not significant
nt	nucleotide
ori	origin of replication
ORF	open reading frame
p53	tumor protein p53
PBS	primer binding site
PCR	polymerase chain reaction
PE	paired-end
PGR	progesterone receptor
PLZF	promyelocytic leukemia zinc finger
Pol2	RNA polymerase II
poly A	polyadenine

PPT	polypurine tract
PPTC7	protein phosphate PTC7 homolog
PR	progesterone receptor
PR	protease
pRB	retinoblastoma protein
Pro	protease
Py	pyrimidine
QC	quality control
qPCR	quantitative PCR
R	repeat region of LTR
R^2	coefficient of determination
RefSeq	NCBI Reference Sequence Database
RFX3	regulatory factor X, 3
RIF1	replication timing regulatory factor 1
RLU	relative light units
Rluc	*Renilla* luciferase
RNA	ribonucleic acid
RNA-Seq	RNA sequencing
RNase H	ribonuclease H
RORA	retinoic acid receptor-related orphan receptor A
RPMI	Roswell Park Memorial Institute medium
rRNA	ribosomal RNA
RSV	Rous sarcoma virus

RT	reverse transcriptase
RT-PCR	reverse transcription polymerase chain reaction
RVLP	retroviral-like particle
Sag	superantigen
SAM	sequence alignment/map
SAS	South Asian
SF1	splicing factor 1
SFFV	spleen focus-forming virus
SINE	short interspersed nuclear elements
SINE-R	SINE of retroviral origin
SIV	simian immunodeficiency virus
SNP	single nucleotide polymorphism
solo LTR	solitary long terminal repeat
Sp1/3	specificity protein 1 and 3
SSBP1	single-stranded DNA-binding protein 1
SU	surface subunit of Env
SV40	Simian virus 40
T	thymine
TAg	T antigen
TAR	trans-activation response
TCGA	The Cancer Genome Atlas
TEF	thyrotroph embryonic factor
TET2	tet methylcytosine dioxygenase 2

TFII	general transcription factor
TM	transmembrane subunit of Env
TNM staging	classification of cancer stage (tumor size, lymph nodes, metastasis)
TRE	Tax response element
tRNA	transfer RNA
trunc	truncated LTR construct
TSS	transcription start site
U3	unique 3' region of LTR
U5	unique 5' region of LTR
UCSC	University of California, Santa Cruz
URI	upper respiratory infection
UV	ultraviolet
v-*onc*	viral oncogene
VCF	variant call format
Vif	viral infectivity factor
Vpr	viral protein R
Vpu	viral protein unique
WDSV	Walleye dermal sarcoma virus
YY1	yin yang 1
ZNF	zinc finger protein
ZNFP	zinc finger pseudogene

Chapter 1: Introduction

1.1 Retroviruses

The *Retroviridae* family consists of a diverse set of enveloped RNA viruses. These viruses, known as retroviruses, are characterized by two distinct traits: the ability to reverse transcribe RNA into DNA as well as the ability to irreversibly integrate their DNA into the genome of their host (1, 2). Since their discovery over 100 years ago, retroviruses have provided unparalleled insight into the building blocks of biology, the mechanisms behind oncogenesis, the evolution of many species, and the host-virus arms race (1-5).

A typical retrovirus virion is ~100 nm wide and encompasses a genome comprised of two dimerized strands of RNA. Each RNA copy is a linear, single-stranded molecule of positive polarity, about 7-11 kb in length (1, 3). The diploidy nature of the retroviral genome can result in a high rate of recombination between related retroviruses and production of heterozygous virions can occur when a cell is infected with multiple viral species (1).

Retroviridae is comprised of seven genera: *Alpha-, Beta-, Gamma-, Delta-,* and *Epsilonretrovirus*, as well as *Lenti-* and *Spumavirus.* The first five genera listed all contain oncogenic species, examples of which include the avian leukosis virus (ALV), mouse mammary tumor virus (MMTV), and human T-lymphotropic virus (HTLV) (1, 2). Lentiviruses, despite not being directly oncogenic, are still greatly pathogenic to their hosts. Human immunodeficiency virus (HIV) is a member of the *Lentivirus* genus that gained notoriety due to its epidemic that surfaced in the 1980s. Its pathogenicity is tied to the deterioration of the host immune system via depletion of CD4-positive T cells,

resulting in acquired immune deficiency syndrome (AIDS) (1, 5-7). To date, retroviruses have been isolated from all mammals and most vertebrate species (3). However, no member of the *Spumavirus* genus is known to be pathogenic (1).

Despite the collectively broad species tropism of the *Retroviridae* family, the method of replication inside the cell is similar amongst most members (**Figure 1-1**). The retroviral life cycle begins with the binding of a virion's surface envelope protein to the host cell receptor and the subsequent fusion of the virion with the plasma membrane (1, 3). A schematic of a retroviral virion is shown in **Figure 1-2A**. The receptors used for entry vary amongst retroviral species and determine the cellular tropism. In the case of HIV, the immune cell-specific CD4 receptor is used (6, 7). Once in the cytoplasm, the viral genome remains encapsidated within the core protein complex. The RNA is reverse transcribed while in this complex to generate a double-stranded DNA copy (1, 2).

Reverse transcription is achieved via the virus-encoded reverse transcriptase (RT) enzyme, a DNA polymerase that the host cell does not normally provide. This process involves the binding of a host tRNA molecule to a string of complementary sequences found towards the 5' end of the viral genome known as the primer binding site (PBS). The polymerase then uses the tRNA as a primer to initiate transcription of the RNA template into DNA (2-4, 8). Reverse transcription was at one point highly controversial, as it defied the central dogma of molecular biology. The central dogma is the principle of only a unidirectional flow of genetic information, whereby DNA is transcribed into RNA that is later translated into protein. This understanding of basic biology was completely reinvented after the discovery of RT (1).

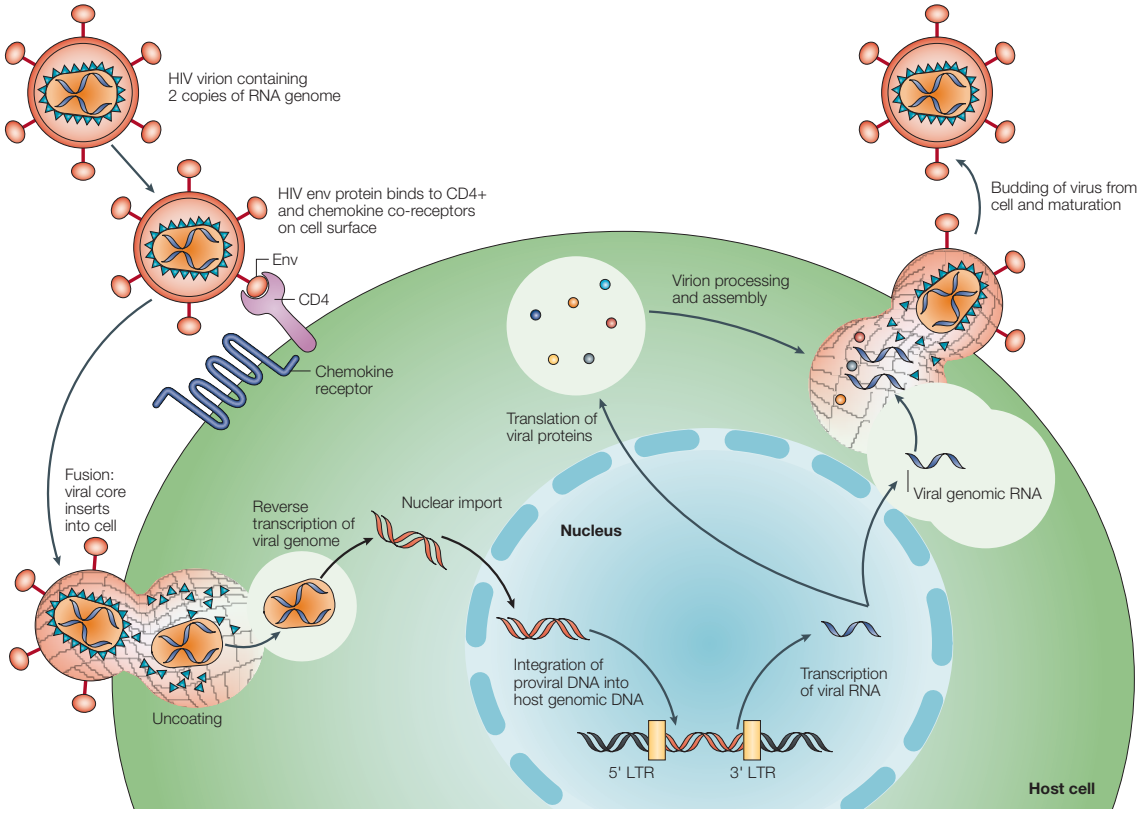


Figure 1-1. Retroviral life cycle.
Schematic of the retroviral life cycle, as exemplified by HIV. Reprinted by permission from Macmillan Publishers Ltd: Nature Reviews Genetics (6), copyright 2004.

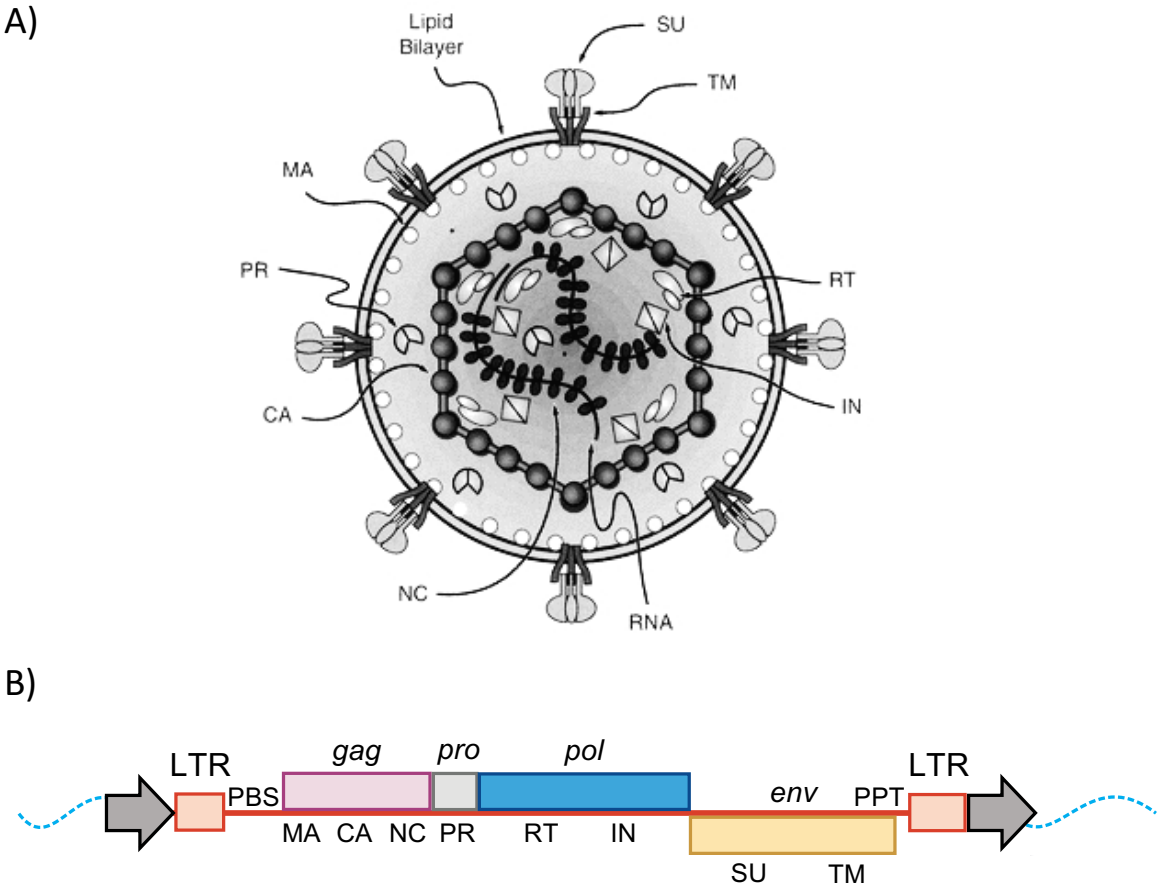


Figure 1-2. Structure of retrovirus virion and provirus sequence.
(**A**) Schematic of the cross-section of a retroviral virion. The envelope consists of a lipid bilayer, stabilized by a series of matrix (MA) proteins. The proteins embedded in the envelope consist of the transmembrane (TM) and surface (SU) proteins. The RNA genome is surrounded by the nucleocapsid (NC) and enclosed by capsid (CA) proteins. Packaged in the virion are the reverse transcriptase (RT), integrase (IN), and protease (PR) enzymes. Reproduced with permission from Cold Spring Harbor Laboratory Press (9). (**B**) Schematic of an integrated simple provirus sequence. Arrows indicate the target site duplications, viral gene names are listed on top, and viral proteins encoded by those genes are listed on the bottom. Reproduced with permission of Annual Reviews in the format Thesis/Dissertation via Copyright Clearance Center (3).

The process of reverse transcription, as shown in **Figure 1-3**, includes two jumps of the reverse transcriptase enzyme, occurring from the 5' to the 3' end of the RNA template. These jumps result in the duplication of the sequences located at each terminus of the template, producing what are known as the long terminal repeats (LTRs) (4, 8). LTRs are critical for driving and terminating viral gene transcription (1, 2, 5). Each LTR is comprised of three unique non-coding regions known as U3, R, and U5. Canonically, transcription is initiated at the U3-R border of the 5' LTR and polyadenylation of the transcript occurs at the R-U5 border of the 3' LTR (3, 4).

After reverse transcription, the viral DNA is translocated to the nucleus and incorporated into the host cell's genome. This random and irreversible process is mediated by the viral integrase enzyme and the integrated viral sequence is known as a provirus (1, 3). A schematic of a proviral sequence is shown in **Figure 1-2B**. The viral genome is flanked by target site duplications, formed by the integration process, as well as the 5' and 3' LTR sequences (3).

Retroviruses can be further classified as either simple or complex, depending on the number of genes they possess (1, 5, 10). Simple retroviruses, as exemplified in **Figure 1-2B**, encode only four genes: *gag, pro, pol,* and *env*. The *gag* gene, found just downstream of the PBS, codes for the matrix, capsid, and nucleocapsid proteins. These are involved with forming and stabilizing the virion as well as viral genome. The three viral enzymes (protease, reverse transcriptase, and integrase) are encoded by the *pro* and *pol* genes. Lastly, the *env* gene encodes the surface and transmembrane components of the envelope glycoprotein (1-4).

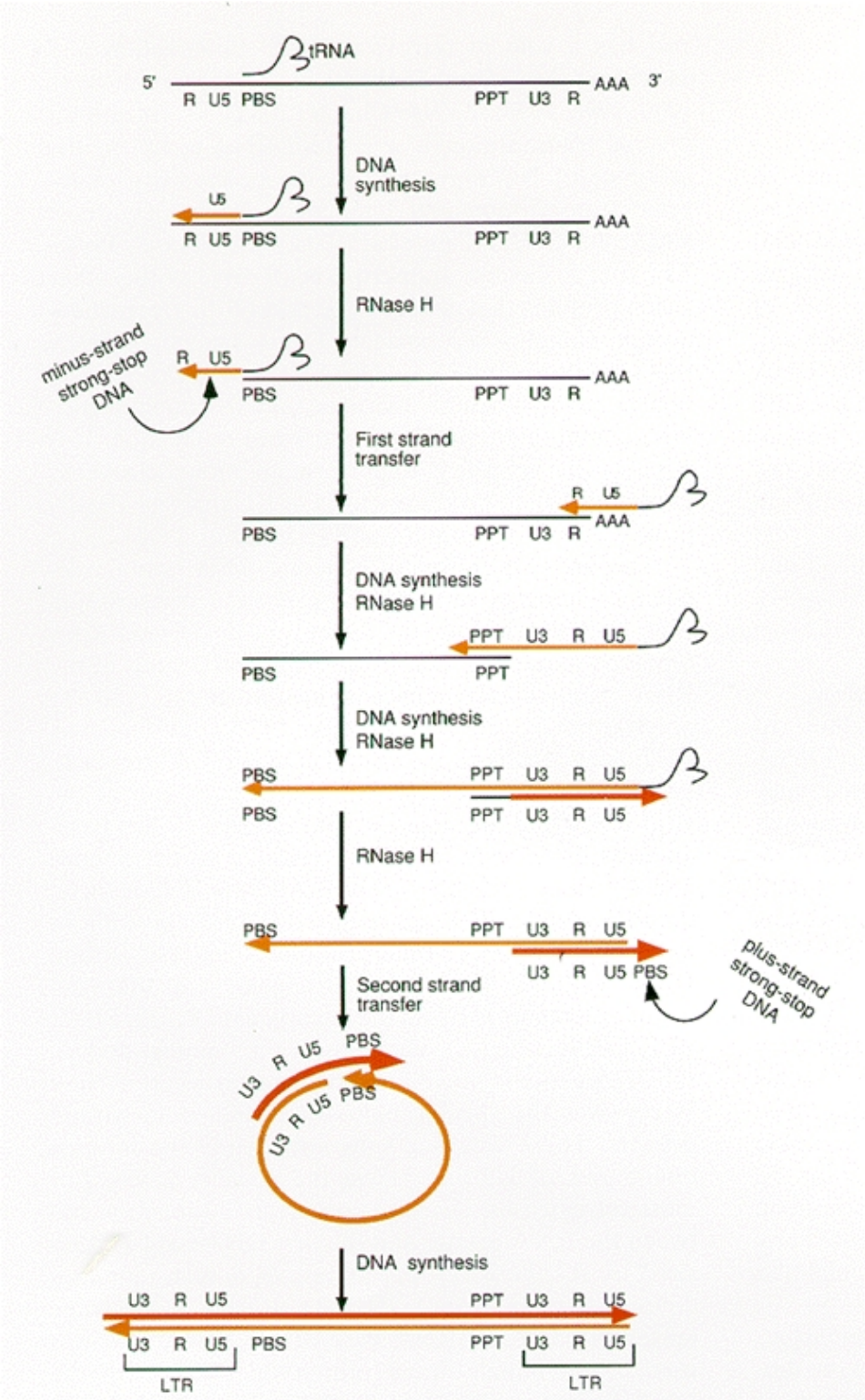


Figure 1-3. Formation of LTR sequences through reverse transcription.
Process of reverse transcription, showing the jump made by the reverse transcriptase enzyme that creates identical LTR sequences at the 5' and 3' ends. Reproduced with permission from Cold Spring Harbor Laboratory Press (8).

HTLV, bovine leukemia virus (BLV), and members of both the *Lenti-* and *Spumavirus* genera are classified as complex. They carry extra sequences known as accessory genes that code for proteins with additional functions that enhance viral gene expression (10-12). Rex and Rev are two accessory proteins that interact with host cell RNA processing machinery to regulate the amount of spliced and unspliced viral transcripts in the cellular cytoplasm. *Rex* is carried by HTLV and BLV, whereas *rev* is part of the HIV genome. Unlike cellular mRNA, not all viral transcripts are spliced in the nucleus; viral structural proteins are translated from unspliced molecules. Rex and Rev have the crucial role of localizing to the host cell nucleus and aiding in the export of unspliced viral transcripts to the cytoplasm (11).

In addition to *rex*, HTLV and BLV encode a second accessory protein known as *tax*. Accessory proteins with functions analogous to Tax are seen in Lentiviruses and Spumaviruses. Many members of these genera carry the *tat* and *bel-1* gene, respectively (11-13). It is important to note that these proteins, while being functionally analogous, are not homologous. Instead, it appears that these accessory proteins have developed independently of one another over the course of evolution (12). Tax, Tat, and Bel-1 operate as transcriptional activators, interacting with host cell factors and the LTR sequence to augment viral transcription (11). Beyond the above-mentioned genes, HIV encodes several other accessory proteins. This list includes Vif, which inhibits the antiviral APOBEC3G enzyme by targeting it for ubiquitination and degradation; Vpr, which is essential for viral replication in quiescent cells as well as importing the pre-integration complex into the nucleus; and Vpu, which enhances the release of the virion

from the plasma membrane as well as prompts the degradation of CD4 inside of the host cell (7, 11).

Once integrated, the provirus is treated akin to any other cellular gene and is transcribed by host cell machinery. For simple retroviruses, transcription is mediated entirely through the interaction of cellular factors with the *cis*-regulatory sequences found in the LTR. For complex retroviruses, regulation of viral gene expression is more advanced due to their *trans*-activating accessory proteins (1, 11). The target sites for these accessory proteins are located in the viral LTR. For example, Tax interacts with cellular factors that bind to the Tax response element (TRE) in the U3 region of HTLV and BLV, Bel-1 binds directly to the Bel-1 response element in the U3 region of the human foamy virus (HFV), and Tat directly binds to the TAR stem loop in the R region of HIV (11, 14).

Provirus transcription results in the production of viral RNAs that are either translated into viral protein or maintained as genomic RNA. Viral proteins and RNA are then assembled at the cell periphery and new viral particles are produced by budding from the plasma membrane. Virion maturation occurs soon after budding and viral polyproteins are cleaved into functional subunits by the protease enzyme (1).

Although most retroviruses follow the life cycle as detailed in **Figure 1-1**, the *Spumavirus* replication strategy deviates in several major ways. In addition to the canonical viral LTR promoter, HFV also encodes an intragenic promoter in the *env* region of its genome that regulates *bel-1* transcription. Bel-1, in return, *trans*-activates this internal promoter, producing a positive feedback loop that controls its own gene expression (11). Spumaviruses are additionally unique in that they undergo the signal-

mediated localization of Env to the endoplasmic reticulum, resulting in glycoprotein maturation, as well as exhibit a high incidence of DNA molecules in some virions (13, 15). Research shows that the majority of HFV virions carry linear double-stranded DNA, suggesting that reverse transcription occurs in the latter half of the life cycle, either during or shortly after budding from the plasma membrane (15).

1.2 Retroviral oncogenesis

Retroviruses were initially discovered due to their oncogenic effects when, in 1908, Wilhelm Ellermann and Oluf Bang isolated an avian erythroblastosis virus from a spontaneous erythroleukemia in a chicken (1, 4). This discovery was quickly followed by Peyton Rous, who noticed the cell-free transmission of a fibrosarcoma in chickens and isolated what was later deemed the Rous sarcoma virus (RSV) (16). By the 1930s, the list of animals exhibiting retroviral-induced tumors had expanded to include cows, cats, mice, and monkeys (1, 17).

As previously mentioned, five of the *Retroviridae* genera include members that are oncogenic (1, 2). These viruses are known to induce various types of tumors in a wide range of species, as detailed in **Table 1-1**. Different mechanisms are used to induce the tumors in these hosts, but in all instances the proviral insertion resulted in the alteration or loss of a sequence required for the proper regulation of a cellular gene (1, 18).

In most cases of retroviral-induced tumors, the impacted sequences are either proto-oncogenes or tumor suppressor genes (1, 18). A proto-oncogene is a cellular gene, also referred to as a c-*onc* gene, that promotes cellular transformation after the induction of gain-of-function mutations to its sequence (1, 4, 17, 19). Mutations in these sequences

Table 1-1. Classification of select retroviruses and their pathogenic effects[a].

Genus	Virus	Host	Pathogenic Result
Alpharetrovirus	Rous sarcoma virus (RSV)	Chicken	Sarcoma
	Avian leukosis virus (ALV)	Chicken	Leukemia/Lymphoma
Betaretrovirus	Mouse mammary tumor virus (MMTV)	Mouse	Mammary carcinoma
	Jaagsiekte sheep retrovirus (JSRV)	Sheep	Pulmonary adenocarcinoma
Deltaretrovirus	Human T-lymphotropic virus (HTLV)	Human	Adult T-cell leukemia/lymphoma
	Bovine leukemia virus (BLV)	Cattle	Leukemia/Lymphoma
Epsilonretrovirus	Walleye dermal sarcoma virus (WDSV)	Fish	Dermal sarcoma
Gammaretrovirus	Murine leukemia virus (MLV)	Mouse	Leukemia/Lymphoma
	Feline leukemia virus (FeLV)	Cat	Leukemia/Lymphoma
	Gibbon ape leukemia virus (GaLV)	Primate	Leukemia/Lymphoma
	Koala retrovirus (KoRV)	Koala	Leukemia/Lymphoma
Lentivirus	Human immunodeficiency virus (HIV)	Human	Acquired immunodeficiency syndrome (AIDS)
	Simian immunodeficiency virus (SIV)	Primate	Simian AIDS
Spumavirus	Human foamy virus (HFV)	Human	No known

[a]From Rous (16), Rosenberg & Jolicoeur (18), Bittner (20), Verwoerd *et al.* (21), Martineau *et al.* (22), Denner (23), Fauci & Desrosiers (7), and Vogt (1).

produce oncogenes and generally result in the stimulation of cell growth and survival (5, 19, 24). Tumor suppressor genes function in an opposite manner. They prevent uncontrolled cell growth by enhancing DNA damage repair pathways, cell death pathways, and activating cell cycle checkpoints (19, 24, 25). These contribute to cellular transformation through loss-of-function mutations, promoting the increase in proliferation that is characteristic of a tumorigenic cell (5, 19, 24).

Oncogenic retroviruses induce transformation through one of three mechanisms: acute, non-acute, or *trans*-acting transformation (26). Acute retroviral transformation requires infection of a retrovirus that carries a viral oncogene (v-*onc*) sequence as part of its genome and results in the LTR-driven overexpression of that gene. These retroviruses transform cells very rapidly, with tumor formation occurring only weeks after infection (26, 27).

Acquisition of a v-*onc* is accomplished through a phenomenon known as oncogene capture. This event, depicted in **Figure 1-4**, occurs when a retrovirus integrates upstream of a cellular proto-oncogene (c-*onc*) in the same transcriptional orientation. Although proviral transcription normally terminates at the 3' LTR, during oncogene capture it continues downstream into the c-*onc* gene, producing a hybrid transcript (4, 27, 28). This read-through transcription is estimated to occur about 15% of the time (29). The hybrid transcript is then packaged alongside a normal viral transcript and, during reverse transcription after infection of a new cell, the two transcripts can recombine and incorporate the proto-oncogene sequence into the viral genome. The newly captured oncogene sequence is known as a v-*onc* gene (1, 4).

Examples of known v-*onc* genes and their associated retroviruses are listed in **Table 1-2**. Since procurement of a v-*onc* comes at the expense of some viral gene sequence, almost all retroviruses carrying these sequences are replication-defective and require co-infection with a "helper virus" that can provide the genes necessary for viral replication (4, 18, 26, 27).

Oncogene capture is a very rare event and most retroviral-induced tumors are from replication-competent viruses that exhibit a long latency period before tumors arise

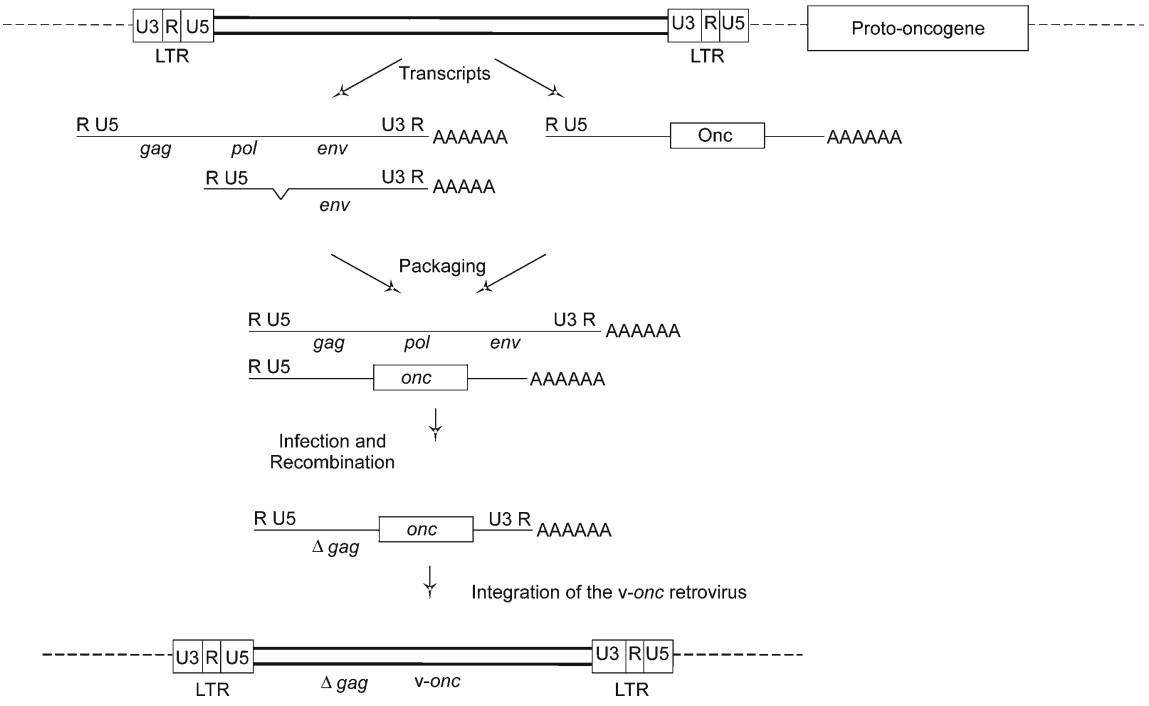


Figure 1-4. Mechanism of oncogene capture.
Schematic of the steps involved with oncogene capture. Illustration shows a provirus integrated upstream of a proto-oncogene and the production of subsequent transcripts: one version is the expected viral mRNA and the other is a hybrid transcript, caused by read-through transcription, containing the proto-oncogene sequence. These two transcripts can be packaged together and undergo recombination, resulting in the acquisition of a v-*onc* gene and loss of some viral sequence. Reproduced with permission of Springer (4).

Table 1-2. Oncogenes overexpressed by retroviral transformation[a].

Mechanism and Representative Viruses	Oncogene(s)
Oncogene Capture	
Rous sarcoma virus	v-*src*
MC29	v-*myc*
Fujinami sarcoma virus	v-*fps*
Avian myeloblastosis virus	v-*myb*
Reticuloendotheliosis virus-T	v-*rel*
Abelson murine leukemia virus	v-*abl*
Moloney murine sarcoma	v-*mos*
Ha/Ki murine sarcoma virus	v-*ras*
Feline sarcoma virus	v-*fes*, v-*fms*, v-*fgr*
Simian sarcoma virus	v-*sis*
Walleye dermal sarcoma virus	rv-cyclin
Insertional Mutagenesis	
Avian leukosis virus	*myc, myb, erbB,* TERT*, bic,* and others
Murine leukemia virus	*myc, myb,* and others
Feline leukemia virus	*myc, myb,* and others
Mouse mammary tumor virus	*wnt-1, fgf,* and others
Accessory Genes	
Bovine leukemia virus	*tax*
Human T-cell leukemia virus	*tax,* HBZ

[a]Reproduced with permission of Springer (27). Only information regarding oncogene capture, insertional mutagenesis, and accessory genes is shown.

(26-28). This mode of transmission, known as non-acute retroviral transformation, is a consequence of insertional mutagenesis. Insertional mutagenesis is the process by which a retrovirus integrates into the vicinity of a c-*onc* gene and the proviral LTR then alters the expression of that gene in a manner that promotes transformation (1, 4, 18). The genes impacted are generally proto-oncogenes that are involved with cellular processes such as mitogenic signaling and cell cycle control (**Table 1-2**) (1, 28).

The means by which the LTR impacts c-*onc* expression depends on its orientation and location with respect to the cellular gene. If the provirus integrates upstream of a

proto-oncogene, and in the same transcriptional orientation, the LTR can directly drive c-*onc* transcription, resulting in aberrantly increased levels (18, 24, 26). If the provirus integrates in the opposite orientation, enhancer elements in the U3 region of the LTR are still able to bind transcriptional activators and stimulate the transcription of nearby genes (26, 27). Enhancer-mediated transcription is the most commonly seen mechanism of insertional mutagenesis. It is effective no matter the orientation and regardless of whether the provirus is up- or downstream of the c-*onc*. Additionally, these sequences are capable of influencing genes at large distances, up to 100 kb away from their site of integration (3, 4, 18).

Proto-oncogene activation is not the only way that retroviral insertions can result in transformation; they can also act through the repression of cellular genes. Insertional mutagenesis can induce upregulation of cellular miRNAs, the targets of which can result in the repression of tumor suppressor genes (28). Additionally, proviral insertion directly into the coding region of a tumor suppressor gene can impact its normal function through the introduction of a pre-mature termination site (28, 30). However, tumor suppressor inactivation is a rare event (27). Since a diploid genome possess two copies of each gene, both need to be damaged before effects can take place. Nevertheless, a cell with one inactivated tumor suppressor copy can become tumorigenic if a second, separate mutation impacts the second copy (18, 19).

Trans-activating transformation occurs independent of proviral integration site and relies upon the expression of certain viral proteins (24, 26). Retroviruses that utilize this method of transformation include MMTV, HTLV, BLV, and certain strains of the murine leukemia virus (MLV). In all cases, the viral proteins interact with host factors in

a way that stimulates the growth of target cells and promotes transformation (4, 11, 12). For HTLV and BLV, the *trans*-activating accessory protein Tax can facilitate the early stages of malignancy. Tax customarily functions to enhance LTR expression by associating with cellular transcription factors on TRE sequences in the U3 region. However, Tax can also activate other cellular genes that play a role in T cell transformation including those that encode the interleukin-2 (IL-2) and IL-2 receptor (IL-2R) proteins. IL-2 and IL-2R are critical for the proper regulation of T cell growth and their overexpression results in an increase in T cell proliferation. It is hypothesized that this increase in proliferation heightens the chances of additional mutational changes in these cells that could contribute to tumorigenesis (4, 11, 31).

Other instances of *trans*-activating transformation involve the binding of viral proteins to host cell surface receptors. The Friend virus (FV) induces transformation through the expression of a mutant Env (4, 32). Canonically, the SU subunit binds to cellular receptors to mediate viral entry (1, 2). However, some mutated SU constructs are able to mimic normal ligand-receptor interactions and activate hormone receptors. This interaction triggers signal transduction pathways associated with host cell growth. FV is a complex of two viruses: the replication-competent Friend murine leukemia virus (Fr-MLV) and the replication-defective spleen focus-forming virus (SFFV). The SFFV virus of the polycythemia-inducing strain of FV (FV-P) encodes for gp55, a defective Env protein. gp55 binds to the erythropoietin hormone receptor, stimulating cell growth and inducing erythroleukemia (4, 32-34).

Similarly, MMTV contributes to malignancy through the binding of superantigen Sag proteins to T cell receptors and major histocompatibility (MHC) class II molecules.

In contrast to other simple retroviruses, MMTV carries an accessory gene called *sag* (4, 35). Exogenous MMTV is transmitted from mother to pup through breast milk and infects B cells. Infected B cells travel to the mammary gland where the virus infects the mammary epithelium, eventually resulting in the formation of mammary gland tumors (20, 35). Sag proteins are able to trigger immense B cell proliferation by binding to and stimulating up to 10% of available T cells (36). This surge in B cell proliferation increases the population of both susceptible and infected B cells, enhancing viral infectivity and cellular transformation (4, 35).

1.3 Endogenous retroviruses

Retroviral transmission primarily occurs horizontally, through infection of somatic cells (**Figure 1-1**). However, in rare instances, infection can occur in the germline of the host and be passed on genetically to future generations. The proviral sequences are then inherited in a Mendelian fashion and are present in the same chromosomal location in every cell of the progeny's body. These congenital integrations are known as endogenous retroviruses (ERVs) (2, 3, 37). To date, retroviruses are the only viruses known to replicate in both endogenous and exogenous forms (3, 38).

ERVs are inherited as novel alleles and, as such, are subject to natural selection. Over time, those with deleterious effects on the host are selected against and ultimately lost from the population. However, those with neutral or advantageous effects can remain, continue to be passed down to future progeny, and ultimately become fixed in the population (1, 3). About 8% of the human genome comprises LTR retrotransposon

sequences, including human endogenous retroviruses (HERVs), that were accumulated in this manner (2, 3, 39).

Currently, there are no known circulating infectious ERV-related viruses in the human population (2, 3, 37). All proviruses in the human genome are the result of bursts of integration events that occurred hundreds of thousands to millions of years ago. The integration of these sequences can be used in the study of species evolution, as their fixed integration sites can be reliably traced through models of evolutionary descent (2, 40). **Figure 1-5** shows a phylogenetic tree of primate species divergence with the approximate integration times of many HERV families. Most HERV families are named using the single letter amino acid code that corresponds to the host tRNA molecule used as a primer for reverse transcription (i.e. HERV-W members use a tryptophan tRNA and HERV-K members use a lysine tRNA) (3, 37).

Phylogenetic analyses of orthologous integrations can be used to determine the approximate age, or time since integration, of ERV sequences. Another method includes a comparison between the 5' and 3' LTRs of a provirus. Canonically, at the time of integration, both LTR sequences are identical (**Figure 1-3**). Over time, mutations accrue and the sequences diverge. The number of nucleotide differences between LTRs, as well as the neutral mutation rate of the host, can be used to approximate age (37, 40).

Using these techniques, the youngest HERV sequences in the human genome are estimated to have integrated between 100,000 and 200,000 years ago (40, 41). Due to the accumulation of mutations over time, no known human provirus still codes for infectious virus (2, 3). The degree of totality exhibited by each provirus varies; some possess only a

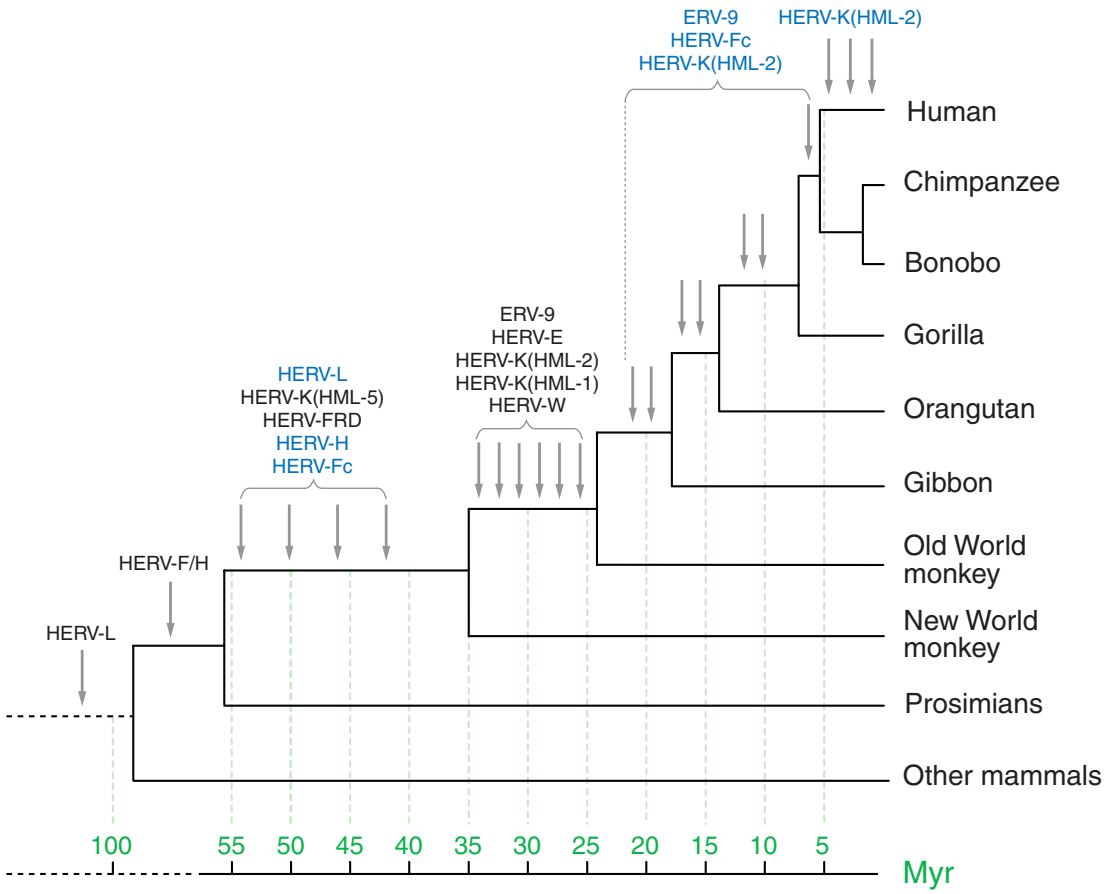


Figure 1-5. Phylogenetic tree of primate species divergence with approximate HERV integration times.
Phylogenetic tree of primate species divergence, with other mammals as the outgroup. Approximate integration times of select HERV families are shown with arrows. Blue font signifies important amplification periods of those HERV families. A timeline is shown in green on the bottom, with years listed as million years ago (myr). Republished with permission of Annual Reviews; permission conveyed through Copyright Clearance Center, Inc. (2).

scattering of point mutations while others suffer from large deletions (40). The most frequent mutation seen is a recombination between the two LTRs, resulting in excision of the internal proviral sequence, leaving a solitary LTR behind (**Figure 1-6**). These elements are known as solo LTRs and they are present in the human genome at a rate 10-fold higher than full- or nearly full-length proviruses (3, 37, 39).

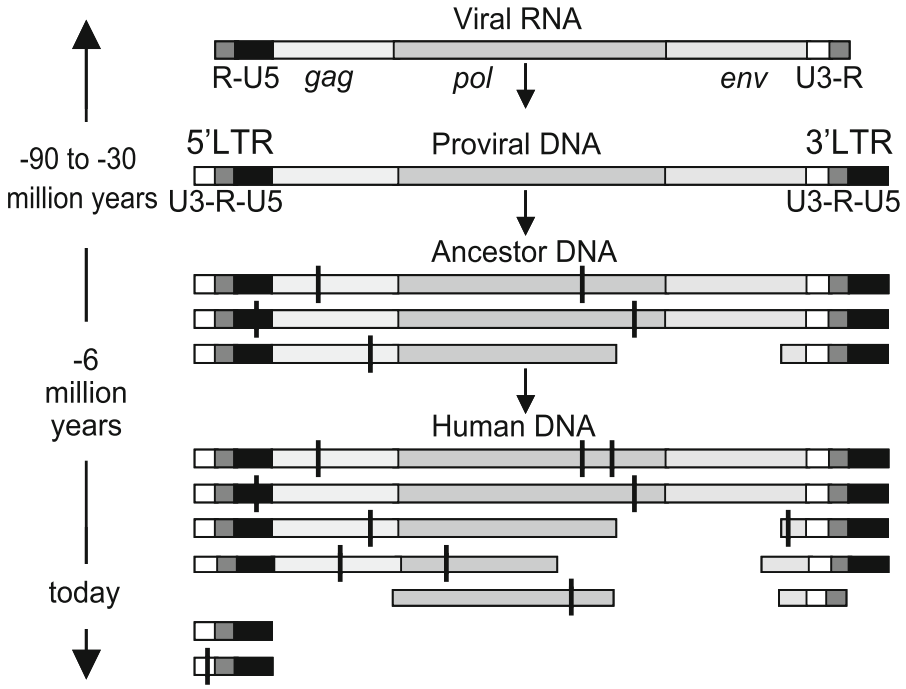


Figure 1-6. Degeneration of HERV sequences over time.
Depiction of the degeneration of HERV sequences over time, beginning with the viral RNA sequence that is integrated as proviral DNA. Mutations are shown with vertical black lines and deletions are shown as gaps in the ancestor and human DNA. Solo LTRs are shown on the bottom. Reproduced with permission of Springer (42).

1.4 Effect of HERVs on host gene regulation

Humans today may not have any circulating infectious ERV-related viruses in their population, but the same cannot be said of other species. MMTV, JSRV, and KoRV are all examples of viruses that currently exist as closely related exogenous and endogenous species. Similar to their exogenous counterparts, newly integrated ERV-related viruses are able to directly alter host gene transcription and drive tumorigenesis. Although the abovementioned viruses do not possess v-*onc* sequences, they can induce tumor formation through non-acute transformation by using promoter and/or enhancer effects to drive nearby c-*onc* genes (43-45). KoRV is of particular interest as it is a relatively young virus, discovered only within the last couple decades, that is currently undergoing endogenization in the koala population of mainland Australia, providing real-time insights into the process of ERV formation (44, 46, 47).

Due to their age and degeneration of most coding capacity, ERVs are considered akin to physical fossils in the human genome; they exist as remnants of ancient integrations from millions of years ago. For many years, these repetitive sequences were incorrectly categorized as "junk DNA" and assumed to play no role in human biology (3, 37, 48). However, with time, research showed that despite being non-infectious, ERV sequences are not necessarily inconsequential. In fact, they are now credited with providing a great degree of genomic plasticity as well as alternative regulation to host gene expression (43, 47, 49-51).

The ability of HERVs to impact host gene expression is extensive. Due to their repetitive nature within the genome, even non-functional sequences are capable of increasing genomic plasticity through homologous recombination. However, a small

number of HERVs are still functional, having retained operative LTRs and open reading frames (ORFs) for viral genes (2, 3). Functional LTRs located near host genes can directly alter cellular transcription in manners similar to those described via non-acute transformation, i.e. the donation of alternative promoters, enhancers, splice sites, and termination signals (2, 52-54). However, HERV expression can influence host transcription through indirect means as well, by upregulating the transcription of regulatory transcripts such as miRNAs and lncRNAs. These non-coding transcripts are known to alter gene expression through RNA silencing and post-transcriptional regulation of certain target genes. Recent work even suggests that non-coding proviral transcripts themselves could function as lncRNAs and play a role in regulating gene expression in human embryonic stem cells (55, 56).

Table 1-3 lists several known instances of HERVs modifying host gene expression. These examples include the tissue-specific effects of a HERV-E sequence that is integrated in the promoter region of AMY1C, resulting in the enhancement of transcription of the amylase digestive enzyme gene in the salivary glands. This enzyme is required for the digestion of dietary starch and glycogen in humans (57, 58). Also noted is a HERV-K integration that provides alternative splice sites for the leptin obesity hormone receptor (LEPR), resulting in additional LEPR isoforms (59).

As a result of natural selection, certain ERV genes have evolved and developed new functions that increase the fitness of the organism through a process known as co-option (60, 61). There are a small number of co-option events that have occurred between ERVs and their hosts. The most well-studied example involves the *syncytin* genes, which

Table 1-3. Examples of HERV effects on host gene expression[a].

Function	Provirus/ Solo LTR	Examples	Reference
Promoter	ERV9 LTR	ZNF80 (zinc finger protein)	(62)
Alternative promoter	HERV-E	APOCI (apolipoprotein CI)	(63)
Bidirectional promoter	HERV-L	DSCR4 and DSCR8 (Down syndrome critical region)	(64)
Promoter, intergenic splicing	HERV-H	PLA2L (phospholipase A2-like)	(65)
Tissue-specific alternative promoter	HERV-P	NAIP (neuronal apoptosis inhibitory protein)	(66)
Promoter	ERV3/HERV-R	H-PLK (human provirus linked Krüppel gene)	(67)
Promoter, enhancer	ERV9 LTR	β-globin locus	(68)
Tissue-specific enhancer	HERV-E	Amy1 (salivary amylase)	(57)
Tissue-specific regulation	HERV-E	PTN (pleiotropin)	(69)
Poly-A	HERV-K (HML-2) LTR	LEPR (leptin receptor)	(59)

[a]Reproduced with permission of Annual Reviews in the format Thesis/Dissertation via Copyright Clearance Center (3). Only information regarding function, provirus/solo LTR, examples, and references for ERVs are shown. Reference numbers are updated to reflect the bibliography of this document.

are formed from preserved full-length ERV *env* sequences and are expressed in the placenta of eutherian mammals. The placenta functions to provide nutrients and oxygen to the fetus while *in utero*. Critical to proper placenta development is the formation of the syncytiotrophoblast layer, which facilitates the maternal-fetal exchange of blood. This exchange is vital for the regulation of oxygen, temperature, hormones, waste, and nutrients, as well as protecting the fetus from rejection by the mother's immune system (70, 71).

Syncytins employ fusogenic mechanisms, similar to those used during *env*-mediated viral entry, to orchestrate the cell-cell fusion of trophoblasts during syncytiotrophoblast production (71-73). Interestingly, co-option of *syncytin* genes for placenta morphogenesis has occurred at least seven times during mammalian evolution. Each of these co-option events has been independent of one other, signifying that their functions arose through convergent evolution (74-79). The human genome codes for two Syncytin proteins: Syncytin-1, encoded by a HERV-W *env* gene on chromosome 7, and Synctyin-2, encoded by a HERV-FRD *env* gene on chromosome 6. These two genes are estimated to have integrated into the primate genome about 30 and 45 million years ago, respectively (70, 72).

Another recently described example of HERV co-option involves the expression of HERV-K loci during human embryogenesis. HERV-K transcription was found to be upregulated in human blastocysts and proposed to play a role in inducing viral restriction pathways to provide immunoprotection of the embryo against exogenous viral infections (80). The immunoprotective potential of ERVs has been documented before in other species. Mice possess a restriction factor known as Fv1 that is derived from the *gag* sequence of a MERV-L element. Fv1 is believed to restrict infection of some retroviruses, including many different strains of MLV. Research suggests that it functions by binding to the capsid proteins of a recently entered virion and preventing the viral DNA from entering the nucleus (81-83).

Other ERVs are able to restrict viral infection through the mechanism of receptor blocking. This was first observed by ALV proviruses, which encode Env glycoproteins that bind to host cell receptors used for viral entry and prevent further infection by

exogenous ALVs (84, 85). Similar methods are utilized by ERVs in mice and sheep. Fv4, an *env*-derived murine protein, can block exogenous MLV infection through receptor interference (83). Likewise, Env proteins expressed from endogenous JSRV proviruses can bind to viral receptors utilized by exogenous JSRV, preventing infection by new virions (86).

Despite not having any single infectious provirus in the human genome, the risk of recombination from several loci with functional ORFs to produce an infectious particle is still a possibility. Previous studies propose evidence of recombination between HML-2 proviruses in HIV-infected individuals (87) and demonstrate the spontaneous production of a recombinant infectious ERV chimera in immune deficient mice (88). Additionally, two laboratories have produced weakly infectious virions from HML-2 proviral sequences *in vitro*. One reconstructed sequence, termed *Phoenix*, was derived from portions of only three full-length proviruses (6q14.1, 8p23.1a, and 7p22.1) (89) whereas the sequence of the other, termed HERV-K_{CON}, was produced via HML-2 consensus genomes (90). Due to the threat of recombination by functional ORFs, as well as the probability of aberrant host gene transcription, ERV activity is not regarded as beneficial aside from those few co-option exceptions. Instead, expression of these proviruses is restricted through epigenetics and ERV sequences are generally silenced under most conditions (**Figure 1-7**) (3, 39, 91).

Experiments using the demethylating agent 5-aza-2'-deoxycytidine show that ERV activity is regulated in part via the methylation of CpG dinucleotides found in the LTR (91, 92). Nearly three-quarters of all human promoters have a higher CpG content as compared to the rest of the genome (93, 94). CpG methylation is a reversible process

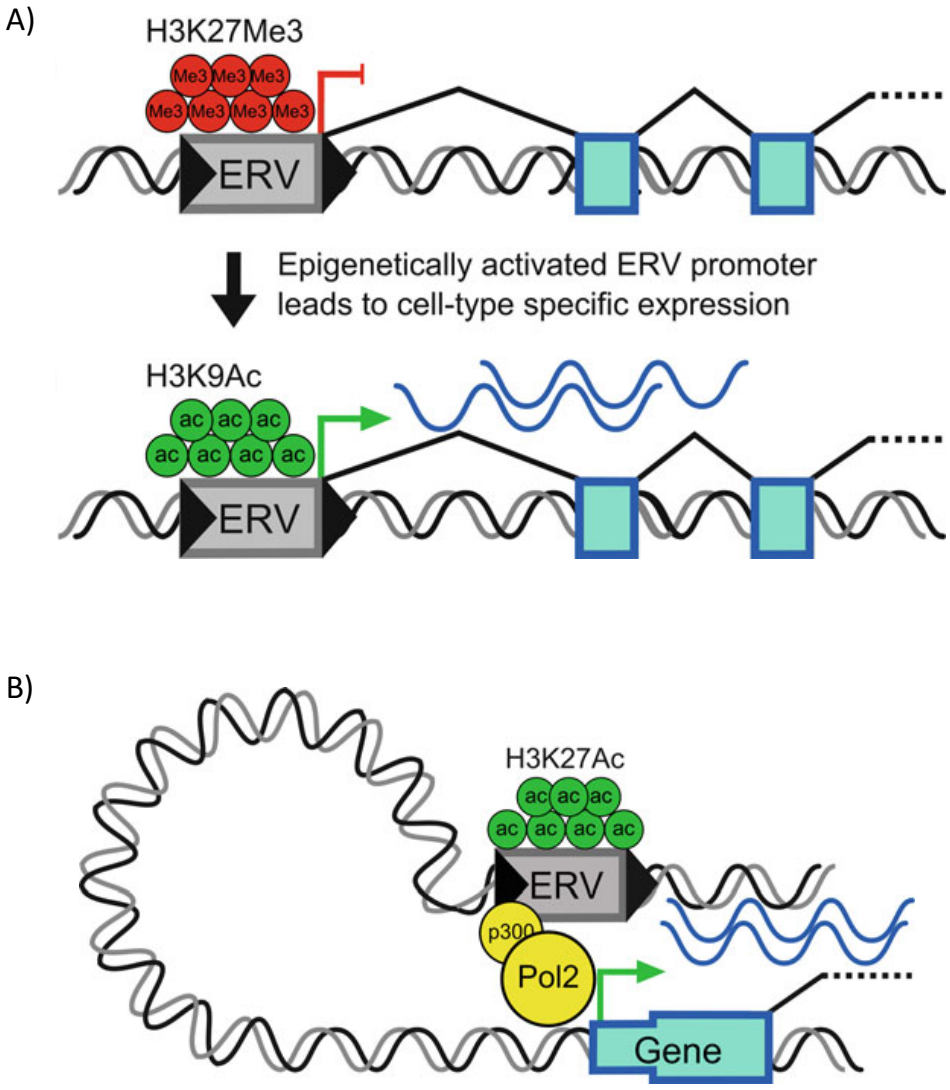


Figure 1-7. Epigenetic regulation of ERV elements.
Schematics showing how epigenetics can regulate ERV transcription. (**A**) Top, repressive histone modifications, i.e. H3K27Me3 methylation, result in silencing of LTRs and prevent gene transcription. Bottom, activating histone modifications, i.e. H3K9Ac acetylation, result in active gene transcription originating from LTR sequences. (**B**) Enhancer-characteristic activating histone modifications, i.e. H3K27Ac acetylation, can also result in gene transcription through enhancer-mediated effects. Reproduced with permission of Springer (95).

whereby a methyl group is added to the carbon-5 position of the dinucleotide's cytosine, forming a 5-methylcytosine (5-mC). Proteins of the methyl-CpG-binding domain family preferentially bind 5-mCs and recruit histone deacetylase (HDAC) and chromatin remodeling complexes. HDACs function by removing acetyl groups found on histone proteins, resulting in more tightly wound DNA. In conjunction with chromatin remodeling complexes, HDAC activity results in a more condensed chromatin state that represses gene activity by directly preventing transcription factors from accessing and binding to the DNA (94, 96, 97). Methylation patterns are established early on in vertebrate development and play a large role in controlling cell type-specific gene expression (39, 54, 95). As such, increased ERV transcription is often restricted to cells that exhibit high levels of hypomethylation, such as those of the placenta, embryonic stem cells, or those associated with certain kinds of disease (3, 92, 98).

1.5 HERV-K (HML-2)

HERV transcripts and subsequent protein products have been detected in both normal and diseased cell types. However, the upregulation of such expression is often associated with malady (2, 99). Members of the HERV-K, -W, and -H groups have all been implicated in various types of human disease, including neurological disorders such as schizophrenia and multiple sclerosis (100-102), autoimmune disorders such as Sjögren's syndrome and systemic lupus erythematosus (101, 103), as well as many types of cancer (104-108). Because HERVs are ubiquitous, any direct role in malignancy is hard to demonstrate. One explanation would be if disease were due to the upregulation of

polymorphic proviruses, but to date no polymorphic HERV element or mutation has been connecteded with a specific disease phenotype (99, 109).

The group most frequently associated with disease is HERV-K (HML-2), which contains the youngest and most well-preserved HERV loci in the human genome. Members of this group began integrating into the primate lineage about 30 million years ago, soon after the evolutionary divergence of New and Old World monkeys. Bursts of HML-2 integration events continued until as recently as 100,000 years ago, resulting in several human-specific integrations (**Figure 1-5**) (2, 41). Overall, the human genome contains over 90 "full-length" HML-2 proviruses and over 900 solo LTRs (40, 41, 110). Despite many possessing large deletions, all full-length proviruses are characterized as those with at least some of their internal proviral sequence intact. Most, but not all, full-length proviruses still possess both LTRs (40). Only infectious retroviruses are officially named by the International Committee on Taxonomy of Viruses and thus HERV sequences currently suffer from a lack of unified nomenclature (111). The most straight-forward method, which will be used throughout this document, is to classify each member by its human chromosomal location as determined by the GRCh37/hg19 build of the human reference genome (40, 112).

The scientific interest in HML-2 proviruses is due predominantly to their young age. Many members still possess intact ORFs, resulting in viral protein production and even retroviral-like particle (RVLP) formation. RVLPs resemble retroviruses, but they are not infectious and do not carry viral genetic material. Instead, they are due to expression of retroviral structural proteins which, when the appropriate concentration is reached, can self-assemble into a virus-like particle (113). RVLPs have been documented

to bud from the cellular membranes of the placenta (114), germ cell tumors (108, 115), melanoma (106), and breast cancer samples (116).

A subset of HML-2 sequences is found only in humans. Human-specificity is due mostly to novel human-specific integrations that occurred after the evolutionary split between humans and chimpanzees (**Figure 1-5**) but also to segmental duplications (2, 40). Segmental duplications are large stretches of identical genomic sequence that are found in different chromosomal locations. Commonly, the novel integration occurred in the genome of an ancestral primate, but the duplication event only occurred within the human population, resulting in a human-specific proviral sequence (117, 118).

Many human-specific integrations are also polymorphic within the current population. About 10% of all HML-2 proviruses are insertionally polymorphic; some individuals have a pre-integration site and no viral sequence, whereas others carry a retroviral insertion either as a full-length provirus or a solo LTR. Other proviruses are allelically polymorphic; they are present in everyone but not necessarily in the same form (i.e. some individuals carry a full-length provirus while others carry a solo LTR or a tandem repeat) (41, 99). Polymorphic HERV sequences imply recent integration times, since it is estimated to take about 40,000 generations for a neutral HERV to become fixed (2, 3).

1.6 Upregulation of HML-2 in breast cancer

HERV-K (HML-2) proviruses were initially detected due to their sequence similarity to MMTV; its name being an abbreviation for Human MMTV-like, group 2. This discovery resulted in the speculation of a causal role for HML-2 in human breast

cancer development (20). It is estimated that nearly a third of all cancer diagnoses in American women are for mammary carcinomas and that the odds of a woman developing breast cancer at some point in her lifetime are 1 in 8. Although some have strong genetic ties, such as those with BRCA1/2 mutations, mammary tumors with known hereditary risk factors only make up 5-10% of all cases (119-121). In fact, nearly 85% of all women diagnosed with breast cancer have no family history of the disease, suggesting strong environmental and lifestyle influences (122, 123). Overall, carcinogenesis is believed to be a multi-step, or "multi-hit", process that requires the accumulation of multiple mutations over time (124, 125). This hypothesis is rooted in the observation that the incidence of cancer increases considerably with age and that it often takes decades after an environmental event, such as UV irradiation, to develop a tumor (124).

Despite an estimated 20% of all human cancers being of viral etiology, there is currently no known circulating virus with a causal role in breast cancer tumorigenesis (3, 126, 127). However, recent studies suggest that exposure to the cattle retrovirus BLV or certain strains of the human papilloma virus (HPV) may increase a woman's risk of developing breast cancer (128, 129). Nevertheless, HML-2 elements are still closely associated with malignancy and increased transcript and protein levels are found in many kinds of tumor samples including breast (50, 105, 119, 126, 130), ovarian (131, 132), melanoma (106, 107, 133, 134), leukemia/lymphoma (135, 136), and germ cell tumors (108, 110, 115).

HML-2 expression is increased in up to 85% of all breast cancer samples as compared to nearby non-diseased tissue (105, 120). This upregulation is seen early on in disease, with significantly increased transcription demonstrated in the serum of patients

with ductal carcinoma in situ (DCIS), widely considered to be stage 0 of breast cancer (137). Viral proteins and antibodies targeting these proteins are also found in the serum of patients, suggesting strong immunological responses to the overexpression of HML-2 proviruses (107, 135, 137, 138). Importantly, anti-HML-2 antibodies as well as RT activity are detectable in afflicted patients only and not in non-diseased controls (120, 131, 136).

The role, if any, that HML-2 is playing in breast cancer development is unclear. Studies have found no correlation between polymorphic provirus expression and tumorigenesis nor any significant associations with HML-2 expression and TNM stage, age at diagnosis, tumor histology, or molecular subtype (109, 122). TNM staging is the most widely used classification of solid tumors, describing the size of the primary tumor (T), spread to nearby lymph nodes (N), and metastases to distant organs (M) (139). Molecular subtype of breast cancer tumors is determined by hormone receptor status and generally classified into three categories: luminal, HER2+, and basal-like.

Luminal cells display estrogen receptor (ER) and/or progesterone receptor (PR) on their cell surface and share similar expression markers with the luminal epithelial cells that line non-diseased breast ducts. HER2+ cells do not express ER or PR and instead are characterized by the overexpression of the *ERBB2* proto-oncogene, resulting in the upregulation of human epidermal growth factor receptor 2 (HER2/*neu*). These cells share expression patterns with those of non-diseased breast tissue. Basal-like are also known as triple-negative cells, as they do not express any of the three hormone receptors on the cell surface. Hormone receptor status is a major factor when determining treatment options for patients. In general, luminal tumors are considered to have the best prognosis,

followed by HER2+, and then basal-like. Basal-like tumors are the least well-characterized and, due to their lack of ER, PR, and HER2 receptors, cannot be treated via targeted therapy (140-142).

HML-2 proviruses may not have direct involvement with initiating carcinogenesis but they may still play active, secondary roles in driving transformation or providing immunoprotection. As stated previously, ERV activation can result in altered host gene transcription. Consequences of this include proto-oncogene activation, alternative isoforms, premature termination, alternative splicing, and increased miRNA or lncRNA production, all of which may contribute to cellular transformation if proto-oncogenes or tumor suppressor genes are impacted (3, 18, 24, 26, 53, 59). Viral protein production is also hypothesized to participate. Tumor cells that express HML-2 envelope proteins on their cell surface may obtain a selective advantage through immunoprotection and escape of immune surveillance (107, 115, 134).

Recent investigations support the claim that HML-2 proviruses encode two accessory genes, known as *rec* and *np9*, that may be oncogenic. Full-length HML-2 members are classified as being type 1 or type 2 and are differentiated based on a 292 bp deletion at the *pol-env* gene junction (**Figure 1-8**). Type 2 viruses encode the full *env* sequence and are capable of producing *rec* transcripts. Rec is unrelated but functionally analogous to the HIV Rev accessory protein and promotes nucleocytoplasmic transport of unspliced mRNA. Type 1 viruses possess the large deletion and are missing the splice donor site necessary to create *rec*. As a result, alternative splicing of this region results in *np9* transcripts (2, 143-145).

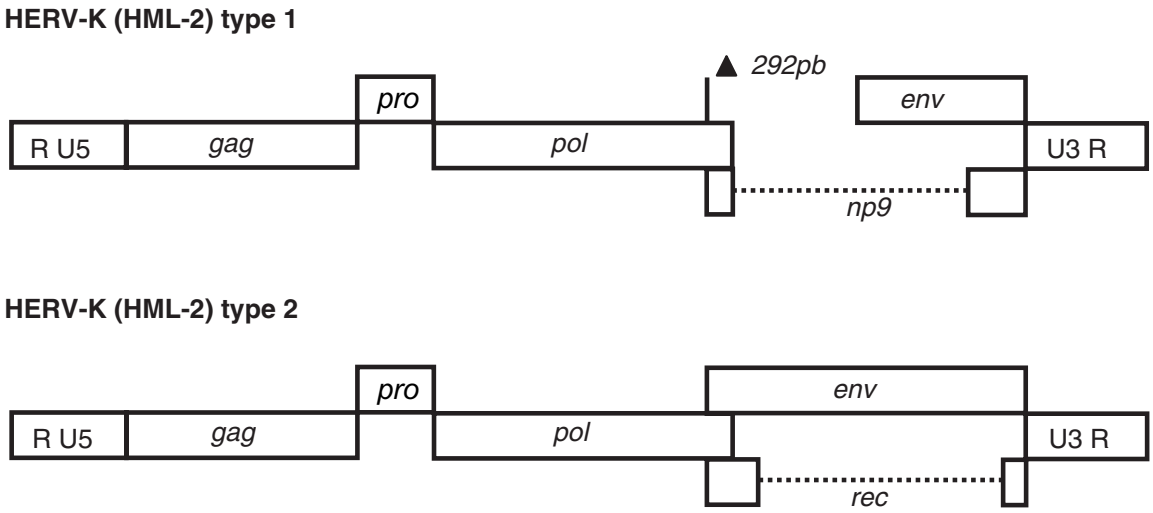


Figure 1-8. Type 1 and type 2 HML-2 viral transcripts.
Genomic organization of type 1 (top) and type 2 (bottom) HML-2 viral transcripts. Type 1 transcripts have a 292 bp deletion at the *pol-env* border and encode the *np9* accessory gene. Type 2 transcripts are full-length and encode the *rec* accessory gene. Adapted with permission of Wiley Materials (143).

Overexpression of Rec and Np9 have been shown to support tumor growth in nude mice as well as interact with promyelocytic leukemia zinc finger (PLZF) protein,resulting in abnormal spermatogenesis and pre-seminoma lesions in transgenic mice (145-147). A major target of PLZF is the c-*myc* promoter and inhibition of PLZF through Rec or Np9 interaction results in de-repression of c-*myc* and an increase in cellular proliferation. This increase in proliferation can aid in the advancement of transformation of a malignant cell (148-150).

The observation that an increase in HML-2 expression is restricted to the sera of cancer patients, and not detected in non-diseased cohorts, suggests a potential role for HML-2 upregulation in immunotherapy or as a molecular biomarker for disease. The

latter is of particular interest as there is currently no simple and accurate way to detect early stages of breast cancer; the current recommendation is for women to be screened regularly through mammograms or self-exams (137). Serum HML-2 mRNA and HERV-K-specific antibody titers were found to correlate with disease onset and reduced patient survival (126, 151), but to drop significantly with cancer treatment (136, 152). These studies suggest that simple blood tests checking for HML-2 activity levels could be used to monitor disease progression and/or treatment efficacy.

Breast cancer patients mount both humoral and cellular immune responses against HML-2 Env-expressing tumor cells, suggesting a possible role for these proteins as tumor-associated antigens and targets for immunotherapy (120, 153). Recent studies show promising results using HERV-K Env-specific chimeric antigen receptor (CAR) T cells to target Env-expressing tumors (154, 155). Others have investigated the use of anti-HERV-K-specific monoclonal antibodies (mAbs). Results show that mice treated with these mAbs have significantly reduced growth of xenograft tumors and that breast cancer cells treated *in vitro* have slowed growth and increased levels of apoptosis (138).

1.7 Provirus-specific expression of HERV-K (HML-2)

Although much work has been conducted looking at the consequences of HML-2 expression and how it may be applied to assist patient care, the mechanisms behind this phenomenon are still largely unclear. Previous investigations have focused solely on detecting HML-2 expression as a group, by utilizing DNA hybridization techniques or RT-PCR with LTR- or gene-specific primers (38, 50, 105). These techniques rely upon consensus sequences and are not stringent enough to distinguish between individual

proviruses, whose sequence similarity can be over 99% (40, 156). They also do not take into consideration proviruses with large deletions or abundant mutations, such as those caused by APOBEC3G proteins, which may no longer possess the consensus sequences used to create the hybridization probes or primers (40, 157). Additionally, unless strand-specific primers are used, these methods cannot distinguish orientation of transcription.

Orientation of transcription is an important factor to consider when determining mechanism and consequence of HML-2 expression. Sense transcripts are of positive polarity and can be directly translated into protein (158). The production of these transcripts from HML-2 proviruses, depending on the ORF functionality, can have significant effects on the host. As stated previously, HML-2 protein production can stimulate the host immune response and is being investigated as a potential target for immunotherapy. Additionally, it is possible that recombination between transcripts with functional ORFs could lead to the production of a novel infectious ERV-related virus (3, 120, 153). Antisense transcripts are complementary to sense mRNA and considered to be non-coding. However, although they are unable to produce protein, they can still modify host gene expression. Antisense RNA molecules, including lncRNAs and miRNAs, have been shown to bind to their complementary sequences and quench mRNA expression, consequently leading to an inhibition of protein production (28, 158, 159).

Retroviral transcription is customarily driven by the 5' LTR. Overall, HML-2 LTRs phylogenetically cluster into three subgroups: LTR5Hs, LTR5A, and LTR5B. LTR5B is the most ancestral form, followed by LTR5A which is presumed to have evolved between 10-30 million years ago. Those of the LTR5Hs subgroup (also referred

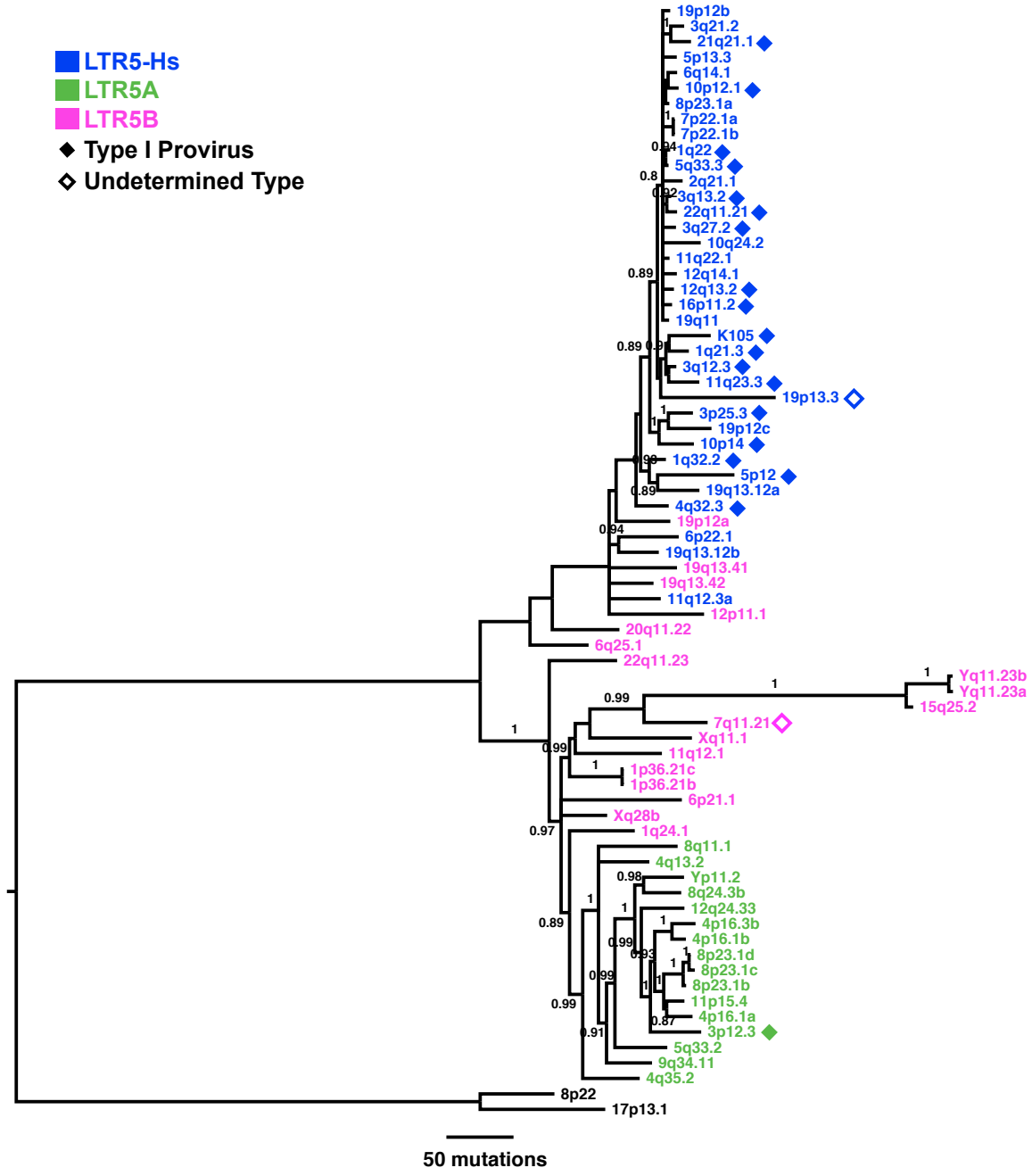


Figure 1-9. Phylogenetic analysis of the HML-2 *env* gene.
Bayesian inference tree of the SU region of *env* of full-length HML-2 proviruses. Provirus names are based on their chromosomal location (hg19 build) and are colored based on LTR subgroup. Proviruses that are type 1 are designated by filled diamonds, proviruses with an undetermined type are designated with open diamonds, and all other proviruses are type 2. Reproduced with permission of BioMed Central (40).

to as LTR-Hs) are the most recently acquired and include all of the human-specific integrations and most type 1 proviruses (**Figure 1-9**) (40). Due to their younger age, as compared to the other LTR subgroups, LTR5Hs sequences are postulated to have fewer mutations and a greater chance for retention of promoter activity.

Eukaryotic, and therefore proviral, transcription is directed via the core promoter as well as nearby *cis*-acting regulatory sequences such as silencers and enhancers (**Figure 1-10**). Although there are no completely universal core promoter elements, most include some combination of a TFIIB recognition element (BRE), TATA box, initiator element (Inr), and downstream promoter element (DPE) (160). Canonically, the transcriptional machinery, which includes RNA polymerase II, is recruited by general transcription factors (TFIIs) to the core promoter and begins transcribing at the transcription start site, frequently encompassed by the Inr element and found 25-30 nt downstream of the TATA box (160, 161). However, analyses of HML-2 core promoters show that LTR-driven transcription is most likely Inr- and TATA-independent, despite some proviruses containing putative Inr and TATA box sequences. Instead, activity appears to be dependent upon specific cellular transcription factors and their respective binding sites (162, 163).

Reliable detection of transcription on a provirus-specific level is important, as many studies suggest that HML-2 expression is highly differential. Evidence shows that the activation of certain proviruses is tissue-specific (110, 147, 151, 163, 164) and that reasons for this are likely to be rooted in the LTR U3 sequences, which contain a wide array of transcription factor binding sites (3, 40). Transcription factors make up ~7% of all protein-coding genes in the human genome and are key factors in controlling gene

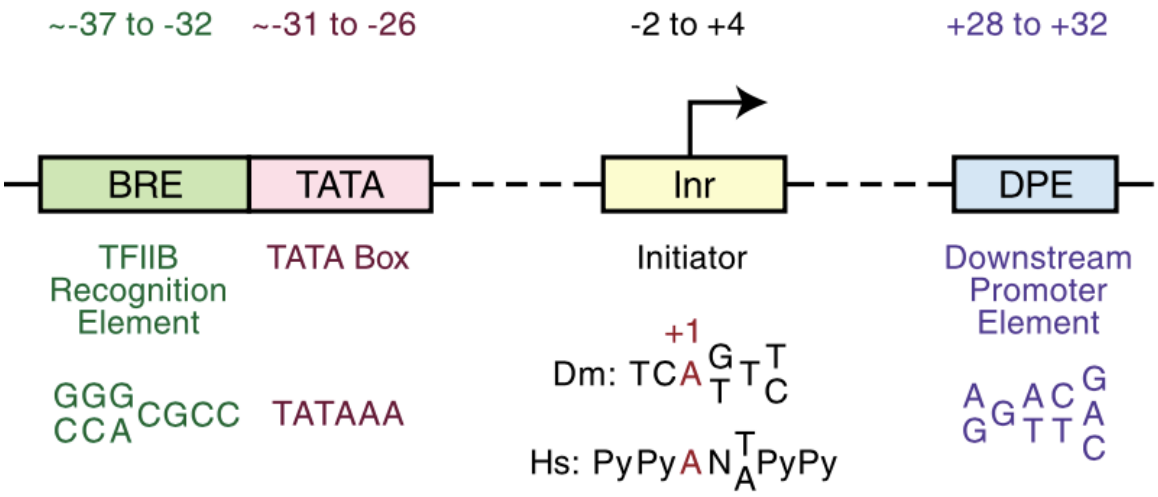


Figure 1-10. Eukaryotic core promoter elements.
Individual core promoter elements are depicted by colored boxes, with corresponding DNA motifs shown underneath and nucleotide position in relation to the TSS shown above. TSS is labeled with a black arrow. Inr motifs are given for *Drosophila* (Dm) and Humans (Hs). Py, pyrimidine. Reproduced with permission from Cold Spring Harbor Laboratory Press (160).

expression and cell type diversity (165, 166). Many transcription factors are cell type-specific and the number expressed in different tissues varies greatly, with about 150 being expressed in the skin to over 300 in the placenta (165). Changes in transcription factor expression is credited with adding to mechanisms of evolutionary adaptation as well as phenotypic diversity within a population (165, 166).

The dysregulation of transcription factor expression is strongly associated with disease, particularly cancer and developmental disorders since most transcription factors orchestrate cellular differentiation and cell cycle control (165, 167, 168). These proteins make up the majority of proto-oncogene sequences and about a third of all developmental disorders are credited with their dysregulation. Investigations into the transcriptome of

leukemia and lymphoma samples show a significant increase in the number of different transcription factors expressed in diseased cells as compared to non-diseased cells (165).

Ubiquitous transcription factors such as Sp1, Sp3, and YY1 are linked with promoting LTR activity but they do not explain the cell type-specific expression that is often seen with HML-2 elements (162, 169, 170). Despite LTR hypomethylation being widely documented to result in promoter activation (92, 98, 171), *in vitro* treatment with DNA methyltransferase inhibitors shows that LTR hypomethylation alone is not sufficient to induce LTR-driven transcription (151, 162, 172), suggesting that the proper transcriptional milieu of a cell is also critical for promoter activation. From these observations, it is hypothesized that cell type-specific transcription factors may play a large role in differential HML-2 expression. However, identification and characterization of these factors have been not been investigated in detail.

Differential HML-2 expression between individuals may also be due to the LTR sequences variation. Mutational differences accrued over time suggest that some LTRs may contain unique binding sites that are responsible for provirus-specific expression. The composition of HML-2 elements in one individual's genome do not necessarily match that of another individual. Overall, the human genome is about 99.5% similar across the species as a whole (173). However, the greatest degree of variance is seen between different ethnic groups and evidence shows that HML-2 sequences can be used to trace patterns of human ancestry (41, 174).

Recently identified non-reference HML-2 elements have estimated insertion frequencies that range from 0.05% to >75% within the human population (41). These frequencies vary between ethnic groups, but the highest degree of variation is found in

African populations. Current theories estimate that anatomically modern humans evolved in Africa about 200,000 years ago and migrated out to the rest of the globe within the past 100,000 years. This migration was followed by a population bottleneck, and non-African populations are believed to have rapidly expanded from these bottlenecks within the past 50,000 years. Because of this, the genetic diversity within African populations is much higher than others, and diversity is hypothesized to decrease the further an ethnic group is from Africa, both geographically and ancestrally (174-176).

Next-generation sequencing (NGS) is becoming the gold standard for analyzing genetic diversity and the cellular transcriptome. RNA-Seq analysis can quantify transcript levels from numerous RNA populations such as mRNA, lncRNA, and miRNA, as well as detect transcription start sites, alternative splice variants, gene fusions, single nucleotide polymorphisms (SNPs), and indels (177, 178). The pipeline involves isolating RNA from cell populations of interest, removing all DNA contamination, reverse transcribing the RNA into cDNA, and then sonicating the cDNA into small fragments. These fragments are PCR amplified, ligated to platform-specific adapters, and sequenced. The resulting sequences are then either computationally *de novo* assembled or aligned against a reference genome to identify the origin of each transcript (178, 179).

NGS techniques have been applied to older groups of HERVs such as HERV-W (180) and –H (181), but they pose some limitations when analyzing younger repetitive elements such as those belonging to HERV-K (HML-2). As stated previously, the sequences of some HML-2 elements are up to 99% similar (40, 156). This degree of similarity creates many difficulties with current RNA-Seq platforms that are dominated by alignments of very short reads, generally 50-100 bp in length (182). These short read

lengths do not allow for differentiation of transcription between similar proviruses. As such, reads of this size that originate from these areas would map to multiple chromosomal locations and not provide definitive quantification of transcription. Additionally, the human reference genome (hg19) that is often used for alignment does not have complete annotation of HML-2 elements. At least 36 HML-2 sequences are known to be either absent or incorrectly annotated in hg19. Non-reference proviruses could be due in part to polymorphisms within the individual(s) used to create the reference or to them being incorrectly thrown out during genome assembly, a common occurrence for repetitive elements (41, 183).

1.8 Rationale for study

Despite elevated transcript expression of HERV-K (HML-2) being seen in many human cancers, the specifics of this phenomenon, including which proviruses are contributing to this expression as well as their mechanism(s) of activation, are unclear. Using a combination of reporter construct assays and high-throughput next-generation sequencing techniques optimized for the capture of HML-2 expression, the goal of this study was to characterize the HML-2 transcriptome and means of LTR activation in an *in vitro* model of human mammary epithelial cell transformation. RNA-Seq results were used to catalog provirus activity, differentiate between sense and antisense expression, characterize modes of HML-2 transcription, and determine if active LTR sequences were impacting nearby host gene expression. A transcription factor binding site prediction algorithm was used to identify unique binding sites that may be responsible for LTR activation during neoplasia. Genomic analyses of these unique sites were conducted using

data from the 1000 Genomes Project and multiple sequence alignments of homologous proviruses in several non-human primate species. Collectively, the results of this study hope to elucidate the molecular mechanism(s) underlying HML-2 promoter activation and transcription, providing new knowledge on the adaptive coevolution of ERVs within their host cells as well as their potential role as a factor in cellular transformation.

Chapter 2: Materials and Methods

2.1 Cell culture

The HME, HMLE-Her2, HMLE-Ras, MCF-10A, SUM149, SUM159, MDA-MB-361, Hcc1419, Hcc1428, and SUM1315 cell lines were generously donated by the Kuperwasser lab at Tufts University and all other cell lines were obtained from ATCC (American Type Culture Collection, Manassas, VA, USA). All cell lines were grown up as per ATTC's recommendations and detailed information regarding culture methods can be found in **Table 2-1**.

2.2 Single-genome sequencing

The single-genome sequencing experiments described below were conducted by a previous postdoctoral fellow in the Coffin lab, Ravi Subramanian. ZR-75-1, MCF-7, T47D, SK-BR-3, Hcc1954, BT20, Hs578T, and MDA-MB-231 cells were grown to 90% confluency. Cellular RNA was extracted and purified using the RNeasy Mini Kit (Qiagen, Valencia, CA, USA, Cat. No. 74104) and DNase treated to remove all DNA contamination (Turbo DNA-*free* Kit, Ambion, Foster City, CA, USA, Cat. No. AM1907). The RNA was reverse transcribed through RT-PCR, as recommended by the manufacturer's protocol and using an oligo(dT) primer (SuperScript III One-Step RT-PCR System, Invitrogen, Carlsbad, CA, USA, Cat. No. 12574-018). The resulting cDNA was then serially diluted down to 1 genome/sample and amplified through PCR using Taq DNA polymerase (Invitrogen, Cat. No. 10342-020). Two forward primers (5'-TTCCTTTACAAAGTTGCGTAAAGC-3' and 5'-GTTGCGTAAAGCCCCCTTAT-3') and one reverse primer (5'-CACAGACACAGTAACAATCTG-3'), all targeting the

Table 2-1. Culture methods for cell lines used.

Base Medium	Supplements	Cell Lines
MEGM[a]	4 mg/ml bovine pituitary extract[a] 1 mg/ml human epidermal growth factor[a] 1 mg/ml insulin[a] 1 mg/ml hydrocortisone[a] 1 mg/ml gentamicin sulfate/amphotericin-B[a] 0.1 mg/ml cholera toxin[b]	HME HMLE-Her2 HMLE-Ras MCF-10A
DMEM[c]	100 µl/ml FBS[d] 10 µl/ml Pen-Strep[e]	MDA-MB-361 MDA-MB-231 MCF-7 BT20 BT474 Hs578T
Ham's F-12[f]	50 µl/ml FBS[d] 1 µg/ml hydrocortisone[g] 5 µg/ml insulin[h]	SUM149 SUM159
Ham's F-12[f]	50 µl/ml FBS[d] 10 ng/ml human epidermal growth factor[i] 5 µg/ml insulin[h]	SUM1315
RPMI 1640[j]	100 µl/ml FBS[d] 10 µl/ml Pen-Strep[e]	Hcc1428 Hcc1419 Hcc1954 ZR-75-1 T47D
McCoy's 5A[k]	100 µl/ml FBS[d] 10 µl/ml Pen-Strep[e]	SK-BR-3
McCoy's 5A[k]	150 µl/ml FBS[d] 10 µl/ml Pen-Strep[e]	Tera-1

[a]MEGM BulletKit, Lonza, Walkersville, MD, USA, Cat. No. CC-3150; [b]Sigma, St. Louis, MO, USA, Cat. No. C8052; [c]Gibco, Carlsbad, CA, USA, Cat. No. 10566-036; [d]Atlanta Biologicals, Norcross, GA, USA, Cat. No. S11195; [e]Gibco, Cat. No. 15140122; [f]Gibco, Cat. No. 31765-035; [g]Sigma, Cat. No. H4001; [h]Sigma, Cat. No. I2643; [i]Sigma, Cat. No. E9644; [j]Gibco, Cat. No. 61870-036; [k]Gibco, Cat. No. 16600-082

HML-2 *env* region, were used in the reaction. The amplified products were purified using the QIAquick Gel Extraction Kit (Qiagen, Cat. No. 28704) and purified samples were sent out for sequencing. The primers used for sequencing were 5'-GACTCCCAGACTATAACCTGTC-3' and 5'-CGAAGCATCAAAAGCCCA-3'. The sequencing results were BLAT searched using the UCSC Genome Browser (112) to identify expressed proviruses.

2.3 Phylogenetic analysis

The DNA sequence of each proviral 5' LTR of interest was obtained from the UCSC Genome Browser's RepeatMasker Track (112, 184). Raw sequences were obtained by entering the chromosomal location of each LTR, as listed in **Table 2-2** and determined by the GRCh37/hg19 build of the human genome, and using the "Get DNA" function. The sequences were then imported as FASTA files into the MEGA (Molecular Evolutionary Genetics Analysis, v6.06) program for alignment using MUSCLE (Multiple Sequence Comparison by Log-Expectation) (185, 186). Phylogeny of aligned sequences was determined by sequence dissimilarity and a neighbor-joining tree was constructed using a p-distance algorithm. Bootstrap values were determined by 1000 replicate tests.

2.4 Dual-luciferase assay

In effort to analyze retention of 5' LTR promoter activity without epigenetic constraints, a series of dual-luciferase assays were developed. The primers for LTR amplification were selected using the Primer3 program (187). Restriction enzyme cleavage sites were added to the 5' end of the primer sequences for proper vector ligation.

Table 2-2. Primers used to amplify 5' LTRs of transfected HML-2 proviruses.

Provirus	Amplified Region (hg19)	Primer Sequences
1q22	chr1: 155604094-155605800	F: 5'-ATTATAGAGCTCCGTTGACTGAGCCATTACCG-3' R: 5'-ATTATAGGTACCTAAAATCCAGCAGCCCAGGA-3'
3q12.3	chr3: 101410385-101412176	F: 5'-ATTATAGGTACCAAGGAGGCTGAGCAGATGAG-3' R: 5'-ATTATTAAGCTTTTCCAGGGGCATCAGAAACT-3'
3q21.2	chr3: 125609085-125610870	F: 5'-ATTATAGGTACCTCACCACCAACCCAGCTAAT-3' R: 5'-ATTATAGAGCTCTCTGGCGGTTGGGTCTTATT-3'
4p16.1b	chr4 9659397-9660991	F: 5'-ATTATAAGATCTCCCTGGATTCCATAAGCAGA-3' R: 5'-ATTATTAAGCTTATAATGGCCCAATCATTCCA-3'
5p13.3	chr5: 30494698-30496674	F: 5'-ATTATAGAGCTCCGGCTCTGCTACATATTCGC-3' R: 5'-ATTATAGGTACCGGACACATACACCCTCCCAA-3'
7p22.1b	chr7: 4638489-4640273	F: 5'-ATTATAGAGCTCCGTTGACTGAGCCATTACCG-3' R: 5'-ATTATAGGTACCAATACCCACAGCACCCAAGA-3'
8p23.1c	chr8: 12082194-12083779	F: 5'-ATTATAGAGCTCAACCATGGGCGGAATTGTTC-3' R: 5'-ATTATAGGTACCAAGAAGTCCACCTGCCTCAA-3'
11p15.4	chr11: 3476909-3478491	F: 5'-ATTATAGAGCTCAACCATGGGCGGAATTGTTC-3' R: 5'-ATTATAGGTACCAAGAAGTCCACCTGCCTCAA-3'
21q21.1	chr21: 19940483-19942138	F: 5'-ATTATAGAGCTCGGTTTAGACTCTGGTGGCCT-3' R: 5'-ATTATAGGTACCTCATGCAGCCTGTAAGTGGA-3'
22q11.21	chr22: 18926085-18927711	F: 5'-ATTATAGGTACCTGCCTCAACCTCCCAAGTAG-3' R: 5'-ATTATAGAGCTCCGGCTCTGCTACATATTCGC-3'

The primers created for amplification of the full-length proviral LTRs of interest in this study are listed in **Table 2-2**. The primers used for amplification of the full-length 22q11.23 provirus, the 22q11.23 LTR-Hs element, and all truncated 22q11.23 LTR-Hs constructs are listed in **Table 2-3**.

Genomic DNA from Tera-1 cells was purified using the DNeasy Blood & Tissue Kit (Qiagen, Cat. No. 69504) and used as the template for PCR amplification with Taq DNA polymerase (Invitrogen, Cat. No. 10342-020). The PCR products were purified using the QIAquick PCR Purification Kit (Qiagen, Cat. No. 28104), digested with the appropriate restriction enzymes (New England Biolabs, Ipswich, MA), and purified with the QIAquick Gel Extraction Kit (Qiagen, Cat. No. 28704). The LTR sequences were then directly ligated using T4 DNA ligase (New England Biolabs, Cat. No. M0202S) into the multiple cloning region of the pGL4.17[*luc2*/Neo] promoter-less firefly luciferase vector (Promega, Madison, WI, USA, Cat. No. E6721), found directly upstream of the bioluminescent *luc2* firefly luciferase gene. Cells were transformed into DH5α competent cells (Thermo Fisher Scientific, Waltham, MA, Cat. No. 18258012) and selected for antibiotic resistance. Positive clones were purified using a Plasmid Maxi Kit (Qiagen, Cat. No. 12162) and all constructs were fully sequenced to check for PCR-induced mutations before transfection.

Cells were seeded at 100,000 cells/well in a 24-well plate and transfected in triplicate. The pGL4 vector was co-transfected alongside a pRL-SV40 internal control *Renilla* luciferase vector (Promega, Cat. No. E2231) at a 30:1 ratio. The internal control vector was under the control of an SV40 promoter. Opti-MEM reduced serum media (Gibco, Cat. No. 31985-070) and Lipofectamine 2000 (Thermo Fisher Scientific, Cat.

Table 2-3. Primers used to amplify HML-2 elements on the 22q11.23 locus.

Element	Amplified Region (hg19)	Primer Sequences
22q11.23 Provirus	chr22: 23879927-23880916	F: 5'-ATTATAGGTACCGGAGAGAGCAGGGGTTTTCT-3' R: 5'-ATTATAGAGCTCATGGCCCCAATGATTCTGGA-3'
22q11.23 LTR-Hs	chr22: 23878249-23879213	F: 5'-ATTATAGGTACCTGTTGTAAGGGGAGCTGGAG-3' R: 5'-ATTATAGAGCTCTCGAGAGTCCCTTCACCCTA-3'
22q11.23 LTR-Hs Trunc847	chr22: 23878249-23879095	F: 5'-ATTATAGGTACCTGGGATGAACTAGAGGACGC-3' R: 5'-ATTATAGAGCTCCGCTCAGCATATGGAGGACC-3'
22q11.23 LTR-Hs Trunc826	chr22: 23878249-23879074	F: 5'-ATTATAGGTACCGGATGAACTAGAGGACGCCC-3' R: 5'-ATTATAGAGCTCGCGCCGCACCGGTCTCTGAG-3'
22q11.23 LTR-Hs Trunc815	chr22: 23878249-30496674	F: 5'-ATTATAGGTACCAGTGCACAGTTCAAAACCCC-3' R: 5'-ATTATAGAGCTCGTCTCTGAGTTCCCTCAGTA-3'
22q11.23 LTR-Hs Trunc805	chr22: 23878249-23879053	F: 5'-ATTATAGGTACCTGGGATGAACTAGAGGACGC-3' R: 5'-ATTATAGAGCTCTCCCTCAGTATTTATTGATC-3'
22q11.23 LTR-Hs Trunc740	chr22: 23878249-23878988	F: 5'-ATTATAGGTACCGCCACTGCCATCTACTAGGA-3' R: 5'-ATTATAGAGCTCGTAATAGTGGGGAGAGGGCC-3'
22q11.23 LTR-Hs Trunc522	chr22: 23878249-23878770	F: 5'-ATTATAGGTACCGCCACTGCCATCTACTAGGA-3' R: 5'-ATTATAGAGCTCCATTCCATTGCCCAGGGATG-3'
22q11.23 LTR-Hs Trunc435	chr22: 23878249-23878683	F: 5'-ATTATAGGTACCGCCACTGCCATCTACTAGGA-3' R: 5'-ATTATAGAGCTCTCAGCACAGACCCTTTACGG-3'

No. 11668-019) were used for the transfections, as recommended by the manufacturer's protocol. The samples were incubated for 48 hours at 37ºC and then lysed for analysis. Luminescence was measured using the dual-luciferase system (Promega, Cat. No. E1910) on a BioTek Synergy HT plate reader using Gen5 Data Analysis Software (BioTek Instruments, Winooski, VT, USA). Empty vector controls as well as non-transfected cells were measured to determine any cell-specific background signal, which was subtracted from the luminescence measurements. LTR promoter activity was calculated as *luc2* activity normalized against that of the internal *Renilla* control signal and quantified as relative light units (RLU). Graphics were produced using Prism 6 (GraphPad Software, La Jolla, CA, USA).

2.5 HML-2 similarity matrices

The sequence of each full-length HML-2 provirus as annotated within the human reference genome (GRCh37/hg19 build) was obtained from the UCSC Genome Browser's RepeatMasker Track (112). DNA sequences were obtained by entering the chromosomal location of each provirus as listed in **Table 2-2** and using the "Get DNA" function. The raw sequences were input into the Clustal Omega program (The European Bioinformatics Institute, Hinxton, Cambridge, UK) to create a multiple sequence alignment using the HHalign algorithm (188, 189). This program conducted pairwise comparisons between alignments to produce a percent sequence identity matrix. An HML-2 percent expression similarity matrix was created by making pairwise comparisons of significant promoter expression patterns in each of the eighteen cell lines used in the dual-luciferase analysis.

2.6 RNA-Seq library preparation

Approximately 1-2 million cells from the Hcc1954, HMLE-Ras, HMLE-Her2, and HME cell lines were used as input for the TRIzol-PureLink RNA Mini Kit system for separate RNA extractions (Ambion, Cat. No. 15596-026 and Cat. No. 1218301A). The RNA samples were treated with 2U DNase (Turbo DNA-*free* kit, Ambion, Cat. No. AM1907) for 1 hour at 37ºC to remove any DNA contamination. Using a protocol previously published by our lab (190), we confirmed that all traces of DNA were removed by qPCR amplification of the TM region of HML-2 *env*.

Purified RNA was used to produce an Illumina RNA-Seq library using the TruSeq Stranded Total RNA Kit with Ribo-Zero Gold (Illumina, San Diego, CA, USA, Cat. No. RS-122-2301). The Ribo-Zero Gold kit removed all cytoplasmic and mitochondrial ribosomal RNA (rRNA) by binding biotinylated capture probes to complementary rRNA sequences. The probes were then bound by magnetic beads and removed with magnets, depleting the samples of rRNA molecules. The remaining RNA was not sheared in order to keep the fragment length as high as possible. Each RNA molecule was then randomly primed and reverse transcribed during first-strand cDNA synthesis. The RNA template was then degraded via RNase H and second-strand cDNA synthesis was achieved using a DNA polymerase. During second-strand cDNA synthesis, dUTP molecules were incorporated instead of dTTP molecules, allowing for differentiation between the two strands. The resulting double-stranded DNA was adenylated on each 3' end, allowing for the specific ligation of thymine-overhanging Illumina sequencing adapters.

DNA fragments were enriched through PCR amplification. The incorporation of dUTP was crucial for creating stranded RNA-Seq libraries. Because DNA polymerases

cannot process through dUTP molecules, amplification of the second strand was inhibited, only the first strand sequence was amplified, and information regarding strand specificity was preserved. Samples were multiplexed so that 25% of the total input of each sequencing lane was occupied by each cell line. Sequencing was done on the MiSeq benchtop sequencer using the MiSeq Reagent Kit v3 (Illumina, Cat. No. MS-102-3001), which facilitates the hybridization of each fragment to the flow cell lawn and addition of fluorescently-tagged nucleotides complementary to the cDNA template. Base calling was determined by wave length emission and fluorescence signal intensity. This sequencing protocol produced approximately 24 million paired-end (PE) reads, all 301 bp in length. Samples were demultiplexed before analysis with CASAVA-1.8.2 (Illumina).

2.7 RNA-Seq analysis

The Trimmomatic program (191) was used to remove Illumina adapters from raw sequencing files and to filter out any reads that did not have an average quality score of at least 25 (signifying about a 0.3% probability of an incorrect base call) (192). The Qualimap program (193) was used to generate a BamQC report to determine the median insert size of our reads. Our data were found to have an average fragment length of ~320 bp for each cell line, suggesting that the PE reads overlapped by an average of 282 bp. Overlapping PE reads were merged together using the FLASh (Fast Length Adjustment of Short Reads) program (194). Overall, 97-99% of reads were merged and the final read lengths used for alignment ranged from 60-592 bp.

Both merged and unmerged reads were utilized to generate alignments using TopHat v2.0.10 and the short-read mapping program Bowtie v2.1.0 (195, 196). Up to two

mismatches per aligned read were permitted. Reads were aligned to two separate reference genomes: the human reference genome (GRCh37/hg19 build) and an HML-2 reference genome comprised of all known HML-2 elements, including those not annotated in hg19 (41, 183, 197). The HML-2 reference genome consists of 1,073 "chromosomes" that represent 96 "full-length" proviruses, 976 solo LTRs, and 1 prototype SINE-R (a SINE element derived from an HML-2 provirus) (198). This collection includes 36 additional polymorphic HML-2 sequences not annotated in the human reference genome (41). Of these, there are four full-length proviruses that are listed as solo LTRs in hg19 as well as five full-length proviruses that are present in hg19 as pre-integration sites (**Table 2-4**) (41, 183). Previous RNA-Seq studies in our lab have shown that alignments to this HML-2 reference genome are able to successfully detect all known proviruses, even those that are recently integrated and highly similar in sequence (110).

Both stranded and unstranded alignments were performed. Data generated from the stranded alignments contained all RNA transcribed in the sense-direction and are designated as such. Data generated from the unstranded alignments contained both antisense- and sense-transcribed RNA and are designated as "Total RNA". About 1-3% of PE reads aligned discordantly and were removed from further analysis. Discordant alignments occur when PE read pairs do not align in the expected manner (i.e. the distance between them is larger than expected or they are in opposite orientation to one another). Concordant alignments were filtered using SAMtools (199) for uniquely aligned reads only, by keeping only reads with a mapping quality score equal to 50, a score that signifies one unique map to the referene genome. Unfiltered reads, which

Table 2-4. Full-length HML-2 proviruses incorrectly annotated in the human reference genome (hg19)[a].

Provirus	Chromosomal Location (hg19)	hg19 Annotation
1p31.1b	chr1:73594980-73595948	Solo LTR
8q24.3c	chr8:146,086,169	Pre-integration site
10p12.1	chr10:27182399-27183380	Solo LTR
12q13.2	chr12:55727215-55728183	Solo LTR
19p12b	chr19:21,841,536	Pre-integration site
19p12d	chr19:22,414,379	Pre-integration site
19p12e	chr19:22,457,244	Pre-integration site
Xq21.33	chrX:93,606,603	Pre-integration site
K105	chrUn_gl000219:175210-176178	Solo LTR

[a]From Turner *et al.* (183) and Wildschutte *et al.* (41).

include sequences present in more than one provirus, are designated as "Unfiltered" whereas reads that passed the filter are designated as "Unique".

The Cuffdiff program was used to generate transcript abundance levels, a process that is done by normalizing read count against the length of the expressed gene as well as the size of the sequencing library (179). These values were reported in units of FPKM (fragments per kilobase of transcript per million mapped reads) and were further normalized across all four cell lines sequenced in order to compare against one another. In analyses that also involved the comparison against Tera-1 FPKM values, a separate Cuffdiff process was run. The Tera-1 data were originally produced by a previous student in our lab, Neeru Bhardwaj, and published as part of her graduate thesis work (110). Transcript expression levels are provided either as normalized FPKM values or as a percent of total HML-2 abundance (determined by dividing the FPKM of one provirus by the sum FPKM of all HML-2 proviruses). Graphics were produced using Prism 6 and

heatmaps were produced using the RStudio Pheatmap pacakage (RStudio, Boston, MA, USA). Jensen-Shannon distance matrices and gene feature plots were produced from Cuffdiff output files using cummeRbund (179). Visualization of the aligned reads to determine mode of transcription was performed using Integrative Genomics Viewer (Broad Institute, Cambridge, MA, USA) (200). Nearby host genes are as annotated by RefSeq on the UCSC Genome Browser (112, 184).

All RNA-Seq data reported here for the Hcc1954, HMLE-Ras, HMLE-Her2, and HME cell lines have been deposited in the NCBI Gene Expression Omnibus database under accession number GSE84275.

2.8 Transcription factor binding site analysis

The full sequence of each 5' LTR of interest was imported into MatInspector, a transcription factor binding site prediction software provided by Genomatix (201). All duplicated binding sites were removed to produce a list containing all predicted binding sites that are unique to each LTR. This program additionally provides information regarding transcription factors that are known to bind to these sites, based on current literature. Transcript abundance levels of those transcription factors in the Hcc1954, HMLE-Ras, HMLE-Her2, and HME cell lines were determined by Cuffdiff analysis of our RNA-Seq results. Heatmaps were produced using the RStudio Pheatmap package.

Consensus sequences of the HOX-PBX, RFX3, ATF, and RORA binding sites were determined through multiple sequence alignments using the MEGA program (185). New reporter constructs containing the consensus (non-active) sites were created through IDT's gBlocks® Gene Fragments synthesis service (Integrated DNA Technologies, Inc.,

Coralville, IA, USA). These fragments were directly cloned into the pGL4[*luc2*/Neo] firefly luciferase vectors and transfected into cell lines as previously described.

Additional MEGA alignments were run to compare the 5' and 3' LTR sequences of each of the nine proviruses of interest. All unique transcription factor binding sites found in only one LTR were regarded as being "acquired" and any unique binding sites found in both LTRs were characterized as "present during insertion". Sites located in the 7p22.1b and 21q21.1 proviruses were excluded from this analysis since they no longer possess intact 3' LTRs (40).

The allele frequencies of each unique binding site within the human population were calculated from the VCF (Variant Call Format) files of 2,504 individuals, as supplied by phase 3 of the 1000 Genomes Project (202). The VCF files were analyzed computationally using VCFtools, by specifying the genomic coordinates (hg19 build) of each site of interest. All sites with an allele frequency of ≥89% were considered to be fixed in the population. All sites that were classified as polymorphic had allele frequencies ≤52%. No binding site that we identified had an allele frequency intermediate of those two thresholds, i.e. calculated to be greater than 53% but less than 89%.

The VCF analysis also provided information regarding the individual SNPs responsible for causing the polymorphic quality of some binding sites. The rs identification number for each SNP was entered into the Ensembl database (The European Bioinformatics Institute and The Wellcome Trust Sanger Institute, Hinxton, Cambridge, UK) for linkage analysis (203). Using the "linkage disequilibrium" function in the database, the degree of linkage between each SNP was determined through pairwise analysis based on population data from the 1000 Genomes Project.

The HOX-PBX site was further analyzed using several non-human primate reference genome sequences as supplied by the UCSC Genome Browser (112). The Denisovan reference genome was obtained from the Denisova High-Coverage Sequence Reads of the Denisova Seq Track. The chimpanzee, gorilla, orangutan, gibbon, and rhesus reference genomes were obtained from the Vertebrate Multiz Alignment & Conservation Track.

Chapter 3: HML-2 5' LTR promoter activity in breast cancer cell lines

3.1 Selection of HML-2 5' LTRs of interest

The HML-2 LTR contains all elements necessary for driving viral transcription. Although identical at the time of integration, the 5' and 3' LTR sequences diverge over evolutionary time as mutations are acquired. In the current human genome, most proviruses have accumulated numerous, unique LTR mutations and it has been suggested that LTR sequence variation could contribute to differential HML-2 expression, particularly through the alteration of transcription factor binding sites (110, 162, 204, 205). Luciferase assays are excellent molecular tools for assessing gene expression patterns. One advantage of this assay is that disruption of expression from endogenous activity of the cell, such as methylation, acetylation, or RNA interference, is removed (206, 207). This assay allows for the direct analysis of how *cis*-acting sequences (i.e. promoters, enhancers, transcription factor binding sites) interact with the transcriptional environment of the cell (207, 208). Through a series of dual-luciferase assays, we sought to evaluate whether LTR sequence identity is correlated with similar promoter expression patterns during breast cancer tumorigenesis.

Preliminary studies into the phenomenon of increased HML-2 expression in breast cancer began in our laboratory with Ravi Subramanian, a former postdoctoral fellow. Ravi used single-genome sequencing (data not published) to detect provirus-specific transcripts from eight human breast cancer cell lines that represented all three molecular subtypes (luminal, HER2+, and basal-like). His objective was to investigate whether proviral expression in these cell lines was diverse or due to transcription originating from only a small number of HML-2 elements.

From Ravi's data, we produced a list of the top ten highest expressing HML-2 proviruses within each cell line tested (**Table 3-1**). The majority of transcripts detected were from the 1q22 and 3q12.3 proviruses, which accounted for 68% of the total expression (39% from 1q22 and 29% from 3q12.3). The sample size for molecular subtype selection was small; only three luminal cell lines, two HER2+ cell lines, and three basal-like cell lines. However, there were no major associations seen between molecular subtype and HML-2 expression, i.e. there was not one subtype with a considerably higher level of HML-2 transcription than another. HML-2 transcripts from luminal cell lines accounted for 29% of the total number, while 39% originated from HER2+ cell lines and 31% from basal-like cell lines.

Table 3-1. HML-2 transcript levels detected through single-genome sequencing in breast cancer cell lines of varying molecular subtype[a].

Provirus	Breast Cancer Molecular Subtype			Total Number of Transcripts
	Luminal[b]	HER2/*neu*[c]	Basal[d]	
1q22	11	30	28	69
3q12.3	20	17	15	52
11p15.4	8	4	1	13
7p22.1	3	7	-	10
21q21.1	2	-	7	9
3q21.2	2	5	1	8
K105	3	1	1	5
16p11.2	3	-	2	5
22q11.21	-	3	-	3
5p13.3	-	3	-	3

[a]Experiments were conducted by Ravi Subramanian, data not published.
[b]Luminal cell lines: ZR-75-1, MCF-7, T47D; [c]HER2/*neu* cell lines: SK-BR-3, Hcc1954; [d]Basal cell lines: BT20, Hs578T, MDA-MB-231

These preliminary data were used to establish a reporter construct assay to investigate whether proviruses of similar LTR sequences exhibit similar expression patterns in human breast cancer cell lines. The proviruses of interest for this study were chosen from **Table 3-1**, with a few exceptions. We included 8p23.1c, a segmental duplication of 11p15.4, and excluded 16p11.2 and K105. The decision for these exclusions was based on the observations that 16p11.2 no longer possess a 5' LTR and K105 exhibited cloning inconsistencies caused by its unique location within the unassembled centromeric region Un_g1000219 (40, 109). In total, nine 5' LTRs were chosen as our sequences of interest for this study and the alternative names and chromosomal locations of these proviruses are listed in **Table 3-2**.

Phylogenetic analysis of the nine LTRs of interest showed that seven of them classified as being of the LTR-Hs subgroup (**Figure 3-1**). This subgroup comprises many of the youngest HML-2 integrations, including all of the human-specific integrations (40, 110, 209). Of the seven LTRs in this subgroup, all but one of them (3q12.3) was

Table 3-2. Transfected HML-2 proviruses with aliases and genomic coordinates[a].

Provirus	Alternative Names	Chromosomal Location (hg19)
1q22	K102, K(C1b), K50a, ERVK-7	chr1:155,596,457-155,605,636
3q12.3	KII, ERVK-5	chr3:101,410,737-101,419,859
3q21.2	KI, ERVK-4	chr3:125,609,302-125,618,439
5p13.3	K104, K50d	chr5:30,486,760-30,496,205
7p22.1b	K108R, ERVK-6	chr7:4,630,561-4,640,031
8p23.1c		chr8:12,073,970-12,083,497
11p15.4	K7	chr11:3,468,656-3,478,209
21q21.1	K60, ERVL-23	chr21:19,933,659-19,941,962
22q11.21	K101, K(C22), ERVK-24	chr22:18,926,187-18,935,361

[a]From Subramanian *et al.* (40) and Bhardwaj *et al.* (110).

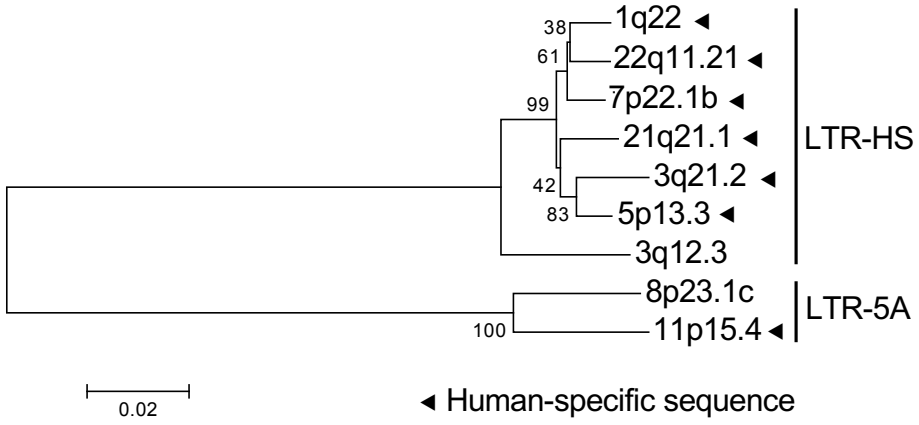


Figure 3-1. Phylogenetic tree of transfected 5' LTRs.
Neighbor-joining tree displaying sequence dissimilarity in the nine HML-2 5' LTRs used in the dual-luciferase assays. Bootstrap values are shown to the left of each node and scale is substitutions/site. LTR type (LTR-HS or LTR-5A) is shown to the right of the tree. Human-specific sequences are designated with a black triangle.

human-specific and phylogenetically clustered close to one another. The 3q12.3 provirus is estimated to be about 5-10 million years old and first integrated into the genome of our last common ancestor with gorillas. The other two proviruses of interest, 8p23.1c and 11p15.4, were of the LTR-5A subgroup. These two proviruses are segmental duplications of each other, but only 11p15.4 is human-specific. It is postulated that 8p23.1c first integrated into the genome of our chimpanzee ancestors about 15-27 million years ago. At some point after the evolutionary split between the human and chimpanzee species, a segmental duplication containing the provirus was copied onto human chromosome 11, resulting in the human-specific 11p15.4 sequence (40).

All nine proviruses are fixed in the human population, although one (7p22.1b) is allelically polymorphic. It is present in the current human population as either a solo LTR or a full-length provirus. However, in either case, the 5' LTR of interest is fixed and as

such, for the purpose of this study, we did not consider any of these nine LTRs to be insertionally polymorphic (40).

3.2 Differential promoter activity of HML-2 5' LTRs in HME cells

The LTR sequences of interest were cloned into a promoter-less firefly luciferase vector (210), directly upstream of the *luc2* gene. The relative promoter activity of these LTRs was determined based on *luc2* expression and normalized against that of an internal control vector, containing a *Renilla* luciferase gene (R*luc*) driven by an SV40 promoter (**Figure 3-2**). These vectors were transiently co-transfected into a panel of eighteen human cell lines. This panel comprised two immortalized human mammary epithelial cell (HMEC) cell lines, fifteen tumorigenic breast cancer cell lines, and one teratocarcinoma

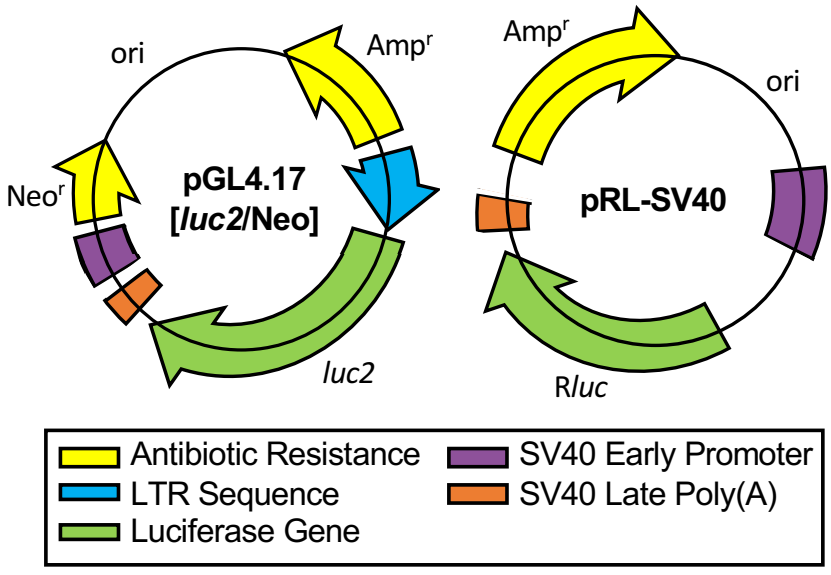


Figure 3-2. Schematics of the reporter constructs used in the dual-luciferase assay. Left, promoter-less firefly luciferase vector (pGL4.17[*luc2*/Neo]). Right, control *Renilla* luciferase vector (pRL-SV40). Direction of gene transcription is shown by arrows. Important gene regions are differentiated by colors and the names associated with those colors are displayed underneath.

cell line (Tera-1). Characterization of the cell lines used for transfection is shown in **Table 3-3**. The Tera-1 cell line was used as a positive control, since it is known to produce high levels of LTR-driven HML-2 transcripts as well as RVLPs and was proposed to exhibit high levels of LTR activity (110, 115). The two immortalized, non-diseased, HME cell lines were used as negative controls, as research suggests that non-transformed cells exhibit little to no HML-2 activity (50, 105, 137).

Overall, negligible levels of promoter activity were detected in the immortalized HMECs while significant upregulation was seen in 73% (11/15) of the tumorigenic breast cancer cell lines (**Figure 3-3**). This expression pattern is consistent with previous reports suggesting that 75-85% of breast cancer samples have a significant increase in HML-2 activity (50, 120, 126). Additionally, each LTR was significantly expressed in at least one cell line tested, but showed differential expression across the panel. Two proviruses

Table 3-3. Characterization of cell lines used for transfection.

Breast Cancer Molecular Subtype	**Hormone Receptor Status**			**Cell Lines**
Luminal	ER+ and/or	PR+	HER2+/-	T47D, MCF-7, Hcc1428, BT474, MDA-MB-361
HER2+	ER-	PR-	HER2+	SUM1315, Hcc1954, Hcc1419
Basal	ER-	PR-	HER2-	MDA-MB-231, Hs578T, BT20, SUM159, SUM149
Additional Cell Types				**Cell Lines**
Immortalized Human Mammary Epithelial Cells				HME, MCF-10A
Tumorigenic Human Mammary Epithelial Cells				HMLE-Her2, HMLE-Ras
Human Teratocarcinoma Cells				Tera-1

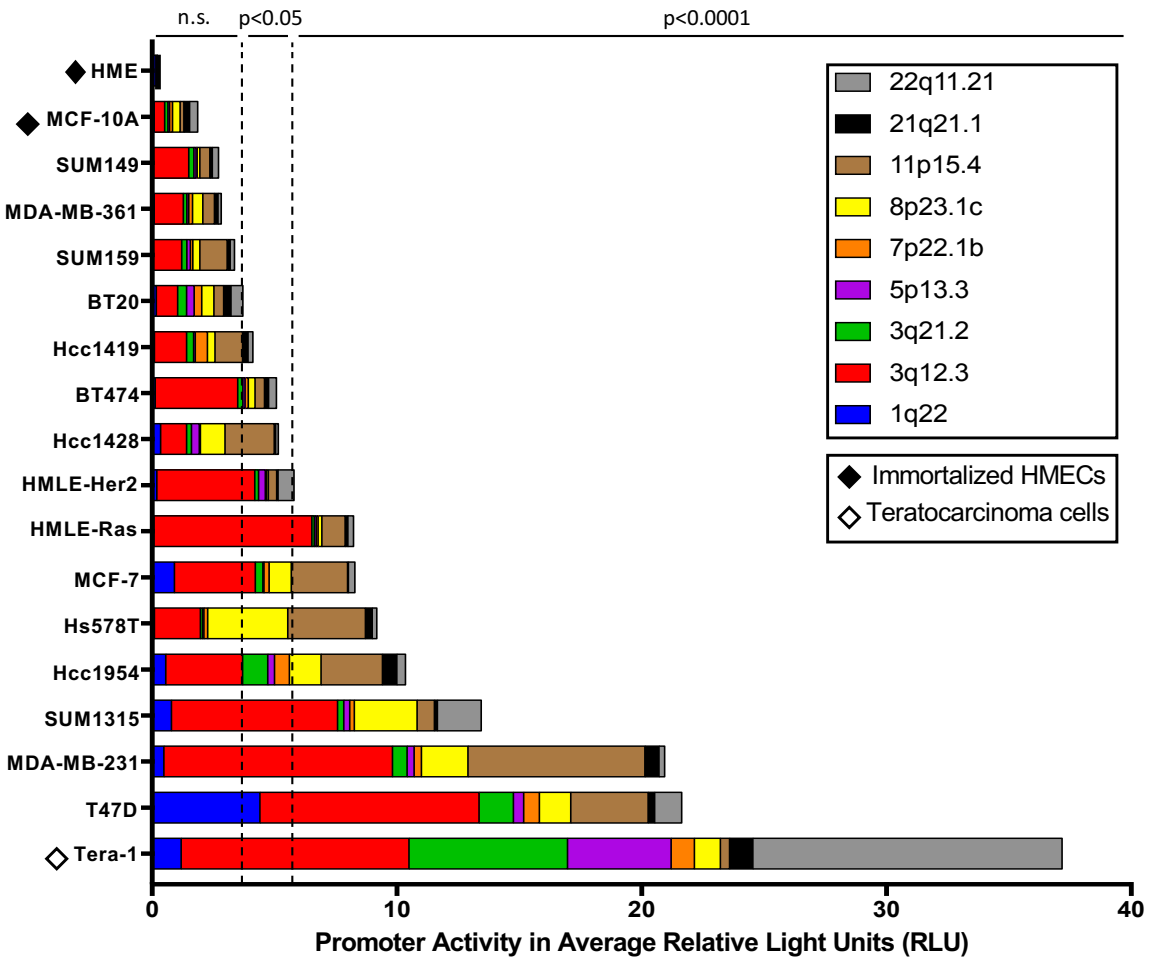


Figure 3-3. Relative 5' LTR promoter activity in eighteen human cell lines. Promoter activity levels are determined as relative light units (RLU) normalized against the internal control *Renilla* expression. Colors delineating each provirus are shown to the right. Immortalized HMECs are designated with a black diamond and teratocarcinoma cell lines are designated with a white diamond. Cell lines without any diamond are tumorigenic breast cancer cell lines. Statistical significance was generated by ANOVA with Bonferroni's multiple comparisons test and is based on comparisons to HME expression (n.s. = not significant) All luciferase experiments were conducted in triplicate and data display the mean ± standard deviation.

(3q12.3, red, and 11p15.4, brown) were significantly upregulated in nearly every neoplastic cell line investigated, whereas others were only upregulated in a select few. Representative images of high, medium, and low expressing proviruses (as exemplified by 3q12.3, 11p15.4, and 1q22 5' LTR promoter activity levels, respectively) are shown in **Figure 3-4**. The 3q12.3 5' LTR was significantly expressed ($p<0.05$) in sixteen cell lines, 11p15.4 in ten cell lines, and 1q22 in three cell lines.

The highest level of combined HML-2 expression was seen in T47D, a tumorigenic cell line known to produce RVLPs under hormonally-stimulated conditions (211-213). However, this activity level was only about half as seen in the Tera-1 cells. Molecular subtype, as denoted by hormone receptor status (**Table 3-3**), showed no significant effect on HML-2 promoter expression as compared to the immortalized HME cell line (**Figure 3-5**). The lack of association between molecular subtype and HML-2 expression is in accordance with Ravi's single-genome sequencing results (**Table 3-1**).

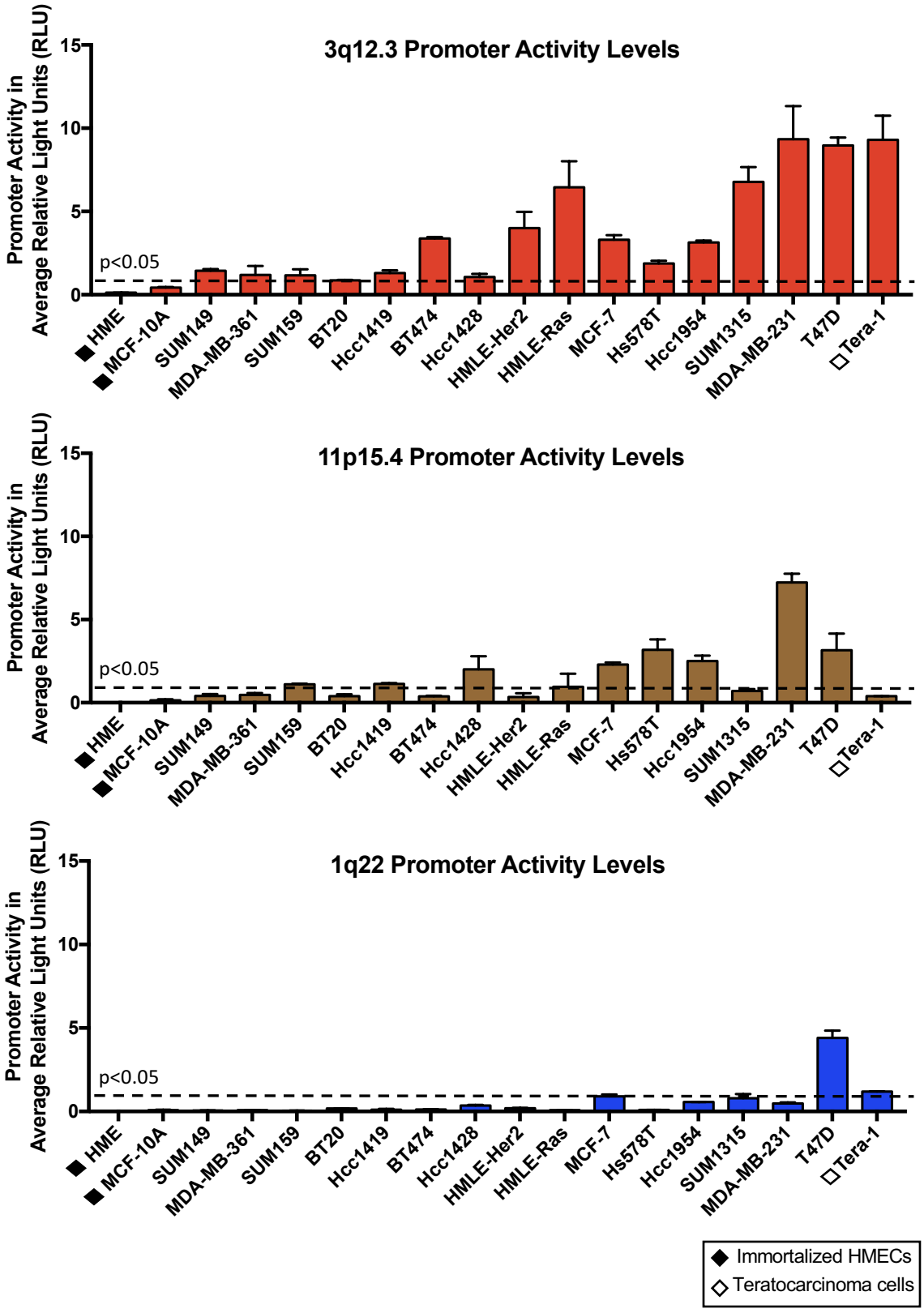


Figure 3-4. Representative high, medium, and low expressing 5' LTR promoters.
Full figure legend is continued on the next page.

Figure 3-4. Representative high, medium, and low expressing 5' LTR promoters. Promoter activity levels are determined as relative light units (RLU) normalized against the internal control *Renilla* expression. 5' LTR promoter activity of the 3q12.3 provirus (top) is shown as representative of a high expressing proviral promoter, the 5' LTR of the 11p15.4 provirus (middle) is shown as representative of a medium expressing proviral promoter, and the 5' LTR of the 1q22 provirus (bottom) is shown as representative of a low expressing proviral promoter. Immortalized HMECs are designated with a black diamond and teratocarcinoma cell lines are designated with a white diamond. Cell lines without any diamond are tumorigenic breast cancer cell lines. Statistical significance was generated by ANOVA with Bonferroni's multiple comparisons test and is based on comparisons to HME expression. All luciferase experiments were conducted in triplicate and data display the mean ± standard deviation.

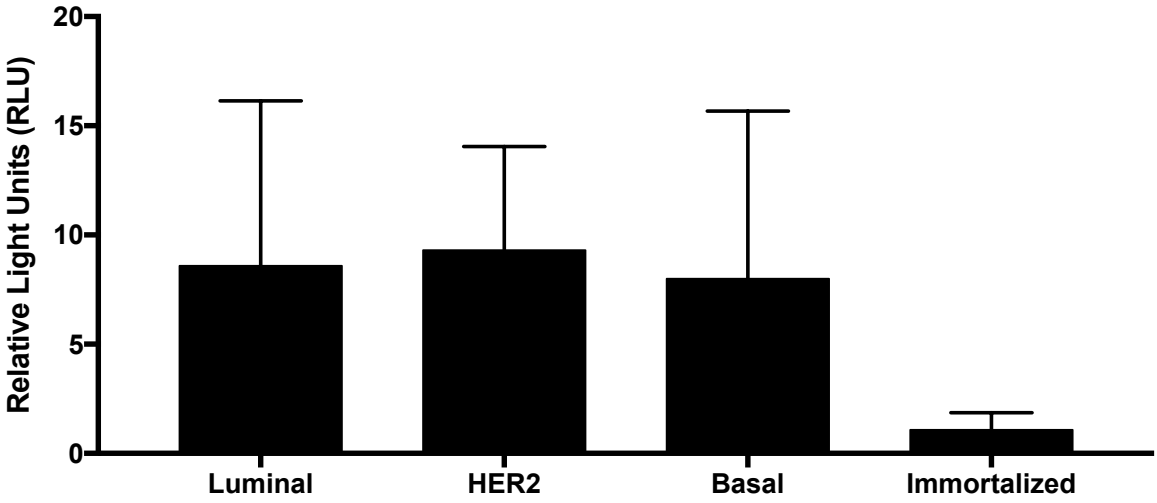


Figure 3-5. HML-2 promoter activity is not breast cancer subtype-specific.
The cumulative 5' LTR promoter activity levels of each of the nine proviruses in this study within breast cancer cell lines of each molecular subtype (luminal, HER2+, and basal-like). These values were compared to promoter activity levels seen in the immortalized HMEC cell lines. Hormone receptor status and cell lines identified as being of each molecular subtype are shown in detail in **Table 3-3**. All values were found to be not significant through ANOVA analysis with Bonferroni's multiple comparisons test based on comparisons to immortalized HMEC expression. All luciferase experiments were conducted in triplicate and data display the mean ± standard deviation.

3.3 Promoter expression patterns are correlated with LTR sequence identity

We next sought to determine if LTRs of similar sequence share similar patterns of promoter activity. For this, we compared values from an HML-2 percent sequence identity matrix, constructed via a multiple sequence alignment using Clustal Omega (188), to an HML-2 percent identity expression similarity matrix, constructed via pairwise comparisons of significant promoter activity within each cell line tested (**Table 3-4**). We found the two values to be correlated, with a coefficient of determination (R^2) value of 0.8133 and a p value of 0.0001, suggesting that LTRs with high sequence similarity are more likely to exhibit significant promoter activity under the transcriptional environment of the same cell line (**Figure 3-6, top**). Overall, LTRs with ~70% sequence identity shared promoter expression patterns ~60% of the time, whereas LTRs with ~95% sequence identity shared promoter expression patterns ~90% of the time (**Figure 3-6B, bottom**). One exception was seen with the 5' LTR of 3q12.3 (**Figure 3-6, red**). This LTR did not demonstrate similar expression patterns to any other LTR. Instead, it exhibited unusually high promoter activity levels, with significant promoter activity seen in almost every tumorigenic cell line investigated (**Figure 3-3**).

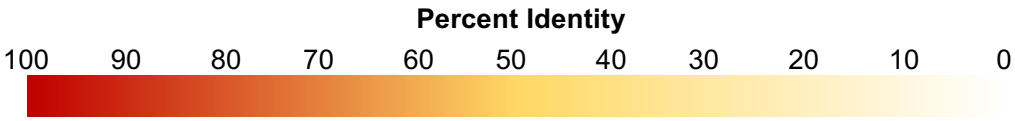
Table 3-4. HML-2 similarity matrices.

HML-2 Percent Sequence Identity Matrix

	1q22	3q12.3	3q21.2	5p13.3	7p22.1b	8p23.1c	11p15.4	21q21.1	22q11.21
1q22	100.00	94.38	97.18	97.16	98.55	69.76	69.02	97.31	98.35
3q12.3		100.00	93.85	94.60	94.48	69.79	68.95	94.68	94.28
3q21.2			100.00	97.15	97.31	68.32	67.93	97.18	97.18
5p13.3				100.00	96.74	69.42	68.45	97.36	96.74
7p22.1b					100.00	69.86	69.13	97.31	98.35
8p23.1c						100.00	95.42	69.17	69.65
11p15.4							100.00	68.74	68.81
21q21.1								100.00	97.31
22q11.21									100.00

HML-2 Percent Expression Similarity Matrix

	1q22	3q12.3	3q21.2	5p13.3	7p22.1b	8p23.1c	11p15.4	21q21.1	22q11.21
1q22	100.00	31.58	94.74	94.74	94.74	68.42	52.63	89.47	89.47
3q12.3		100.00	36.84	26.32	26.32	63.16	68.42	26.32	36.84
3q21.2			100.00	89.47	89.47	73.68	57.89	89.47	89.47
5p13.3				100.00	100.00	63.16	47.37	100.00	89.47
7p22.1b					100.00	63.16	47.37	100.00	89.47
8p23.1c						100.00	73.68	63.16	73.68
11p15.4							100.00	47.37	47.37
21q21.1								100.00	89.47
22q11.21									100.00



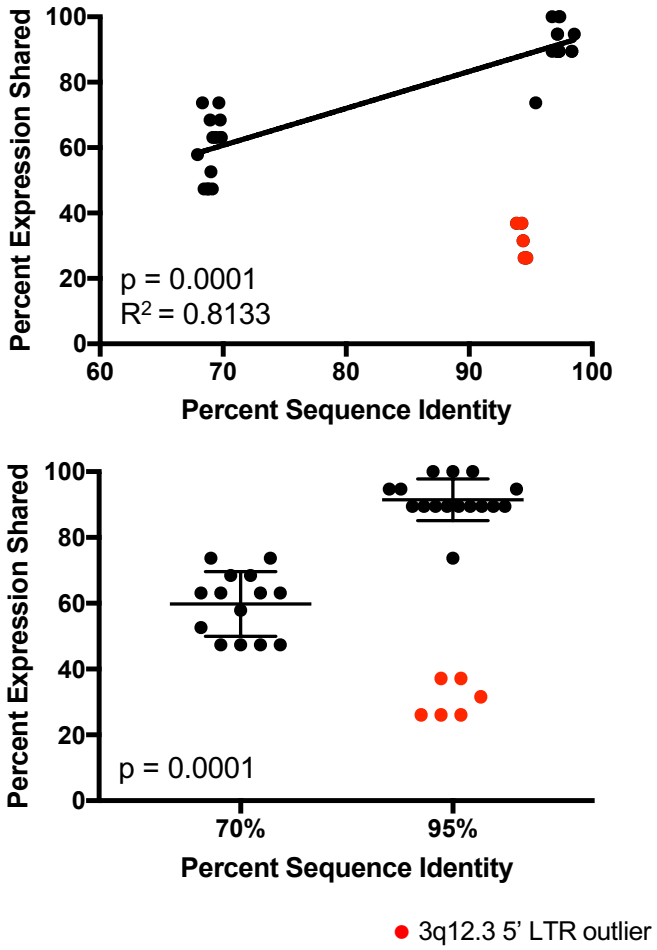


Figure 3-6. LTR sequence identity is correlated with promoter expression patterns, with the exception of 3q12.3.
Scatter plots displaying the correlation between percent sequence identity and percent expression shared. Raw values are shown in **Table 3-4** and are based on pairwise comparison. Best fit line and R^2 values are shown for the top graph. Statistical significance of the top graph was generated by Pearson correlation. Error bars depict the mean ± standard deviation in in the bottom graph. Statistical significance of the bottom graph was generated by unpaired t test with Welch's correction. Outlying 3q12.3 data points are shown in red for both plots.

Chapter 4: Next-generation sequencing analysis of HML-2 provirus transcription

4.1 In vitro model of human mammary epithelial cell transformation

Our reporter construct assays suggest that many HML-2 5' LTRs have retained functional promoter activity that can be activated under the transcriptional environment of a tumorigenic breast cancer cell. However, because these assays were transient transfections, they did not demonstrate what is endogenously occurring within the cell. To assess this, we turned to next-generation sequencing.

Current NGS platforms commonly rely upon the alignment of millions of very short reads, 50-100 bp in length, to the human reference genome (182). These techniques are not adequate for the characterization of HML-2 transcription as HML-2 sequences are often 99% similar to one another and require longer read lengths to differentiate between them (40, 156). Additionally, alignments to the human reference genome (GRCh37/hg19 build) do not allow for detection of all HML-2 elements, as at least 36 are known to be missing or otherwise annotated incorrectly (41, 183).

Alongside a previous graduate student in our lab, Neeru Bhardwaj, I helped produce and validate an NGS protocol that is optimized for the capture of HML-2 expression. This technique, as detailed in the Materials and Methods section of this document, utilized long (301 bp), paired-end (PE) reads, produced from stranded libraries, that were filtered for unique alignments only. PE sequencing allowed for both ends of the RNA molecule to be sequenced and all overlapping PE reads were merged together to produce reads up to 592 bp in length. Stranded libraries detect orientation of transcription, allowing for delineation of sense vs. antisense transcripts. These reads were then aligned against both the human reference genome and a custom HML-2 reference

genome, containing all known HML-2 elements. Only reads that map uniquely, to one chromosomal location, were analyzed. Critically, this method has proven successful for capturing distinct transcripts from young proviruses, with high sequence similarity, as well as proviruses incorrectly annotated in the human reference genome (110).

We chose to use this NGS protocol to characterize which proviruses are expressed during human mammary epithelial cell transformation and to determine the mechanism(s) of HML-2 transcription throughout this process. We utilized an *in vitro* model of HMEC transformation, as originally described by Elenbaas *et al.* (214). This model mirrors the transition of a primary cell to a tumorigenic one in three steps, producing cell lines of varying states of malignancy. This method allows for the analysis of how malignant shifts in the transcriptional environment of a cell may impact HML-2 expression.

A schematic of the transformation model is shown in **Figure 4-1**. HMEC immortalization is achieved through telomere maintenance by *hTERT* (human telomerase reverse transcriptase) overexpression. Telomeres are sections of repetitive nucleotides found at the ends of each chromosome. With each replication cycle, telomeres shorten until they reach a critical length, at which point either cellular senescence or apoptosis is triggered (215, 216). This phenomenon is known as the Hayflick limit and it is believed to prevent genomic instability and protect against cancer by placing a limitation on the number of cellular divisions a single cell can undergo. hTERT is able to lengthen these sequences, allowing them to divide indefinitely. The activity of this enzyme is characteristic of stem cells and tumors, and is imperative for a cancerous cell to achieve immortality (217, 218).

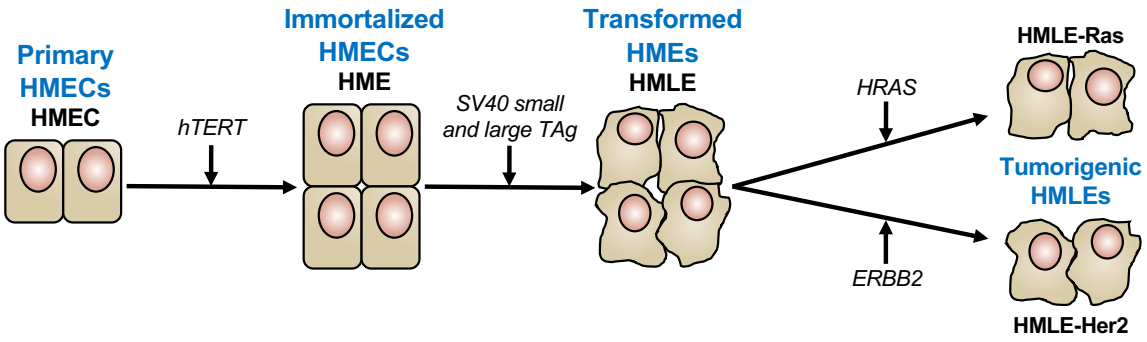


Figure 4-1. Schematic of the HMEC transformation process *in vitro*.
HMEC transformation process, as initially described by Elenbaas *et al.* (214). Transformation steps are shown with black arrows and transformation stages are labeled in blue. Cell lines names and overexpressed genes associated with each step are labeled in black.

Immortalized primary HMECs, known as HME cells, are transformed into HMLE cells via the introduction of the SV40 early region. This region contains the genes encoding the small and large T antigens (TAgs) of the SV40 polyomavirus. These oncoproteins are able to induce transformation of a cell by inhibiting the p53 and pRB pathways, resulting in uncontrolled cellular division (214, 219, 220). Lastly, overexpression of the commonly amplified breast cancer oncogenes *HRAS* and *ERBB2* (also known as HER2/*neu*) is achieved to produce HMLE-Ras and HMLE-Her2 cells, respectively (221-223). The tumorigenic property of the HMLE cell lines is based on observations that they can produce tumors in immune-deficient mice as well as exhibit anchorage-independent growth in soft agar (214, 219).

Validation of the overexpression of the genes used to create these cell lines, as well as characterization of their hormone receptor status, is shown in **Figure 4-2**. The HME cell line (**Figure 4-2A, blue**) had induced overexpression of only *TERT*. The HMLE-Her2 cell line (**Figure 4-2A, red**) had induced overexpression of *ERBB2, TERT,*

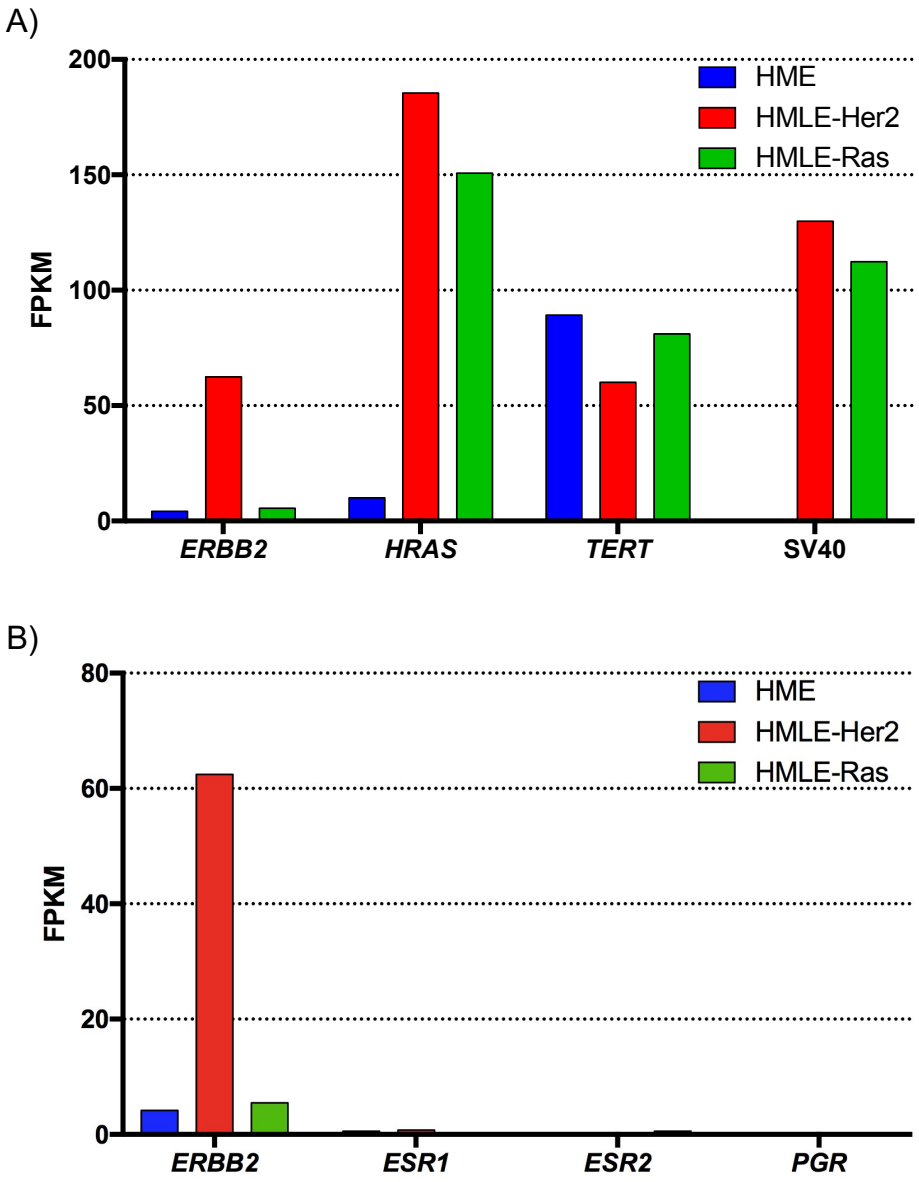


Figure 4-2. Validation of gene overexpression in HME and HMLE cells.
(**A**) Transcript expression levels of genes used to create the HME and HMLE cell lines. SV40, the SV40 early region containing the small and large T antigen genes. (**B**) Transcript expression levels of genes corresponding to the three major hormone receptors in HMECs. All values are given in FPKM as determined by stranded, unique alignments and are normalized across all cell lines using Cuffdiff.

and SV40 early region. The HMLE-Ras cell line (**Figure 4-2A, green**) had induced overexpression of *HRAS, TERT,* and SV40 early region. Since HRAS expression is downstream of the ERBB2 signal transduction pathway (221-223), increased *HRAS* expression was also seen in the HMLE-Her2 cells. All of the cells are negative for the three major hormone receptors with the exception of HMLE-Her2 cells, which overexpress the *ERBB2* gene that encodes the HER2/*neu* receptor (**Figure 4-2B**).

Passage-matched samples of HME, HMLE-Ras, and HMLE-Her2 cells were subject to Illumina MiSeq RNA sequencing alongside Hcc1954, an established tumorigenic breast cancer cell line (224). Statistical analysis using Jensen-Shannon divergence (225), which measures pairwise dissimilarities between multiple conditions, demonstrates the degree of change in global gene expression within each cell line sequenced (**Figure 4-3**). A clear shift in transcriptional environment is seen between the HME cells and their tumorigenic counterparts, suggesting that cellular transformation notably alters the global transcriptional milieu of a cell.

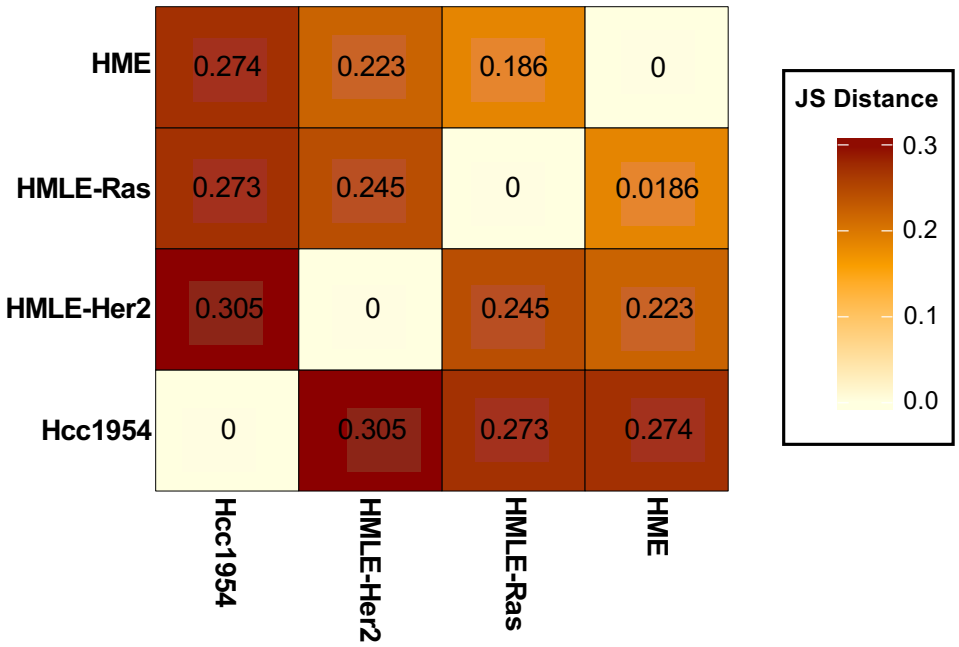


Figure 4-3. Jensen-Shannon divergence matrix of cell lines sequenced. Jensen-Shannon divergence matrix (225), showing statistical pairwise dissimilarity in global gene expression of the four cell lines sequenced. Values are given as JS distance; increased JS distance signifies an increase in dissimilarity. Gene expression values used for the analysis were determined by Cuffdiff.

4.2 Validation of sequencing protocol

We next investigated the HML-2 expression profiles in these four cell lines. To ensure that we were only analyzing results that we could confidently say mapped to a certain provirus, we filtered our data for reads aligning to only a single location in the genome. In this current analysis, 96-99% of our HML-2 reads were retained after filtering and able to align uniquely, suggesting that our protocol is stringent enough to detect all expressed proviruses and to bypass any multi-mapping issues (**Table 4-1**).

Transcript expression levels were calculated as FPKM (fragments per kilobase of transcript per million mapped reads), which takes the number of reads that align to the

Table 4-1. Retention rate of HML-2 reads after filtering for unique alignments only.

Cell Line	Retention of Filtered HML-2 Reads
Hcc1954	99.7%
HMLE-Ras	98.5%
HMLE-Her2	96.7%
HME	96.0%

exons of each transcript and normalizes them against the length of the transcript as well as the total number of reads in the sequencing library. Total HML-2 expression levels were determined by summing the FPKM of each individual provirus. These values were compared against the values from one of our previously published sequencing studies, during which we characterized the HML-2 expression profile in the human teratocarcinoma cell line Tera-1 (110). Since this cell line is known to express HML-2 elements at exceptionally high levels (53, 91, 115), it was used as a proof of concept for our RNA-Seq protocol. In our current investigation, we detected HML-2 RNA in all cell lines sequenced, with the highest levels seen in the tumorigenic Hcc1954 and HMLE-Ras cells. Although these levels were approximately 10-fold lower than those seen in Tera-1 cells, they were comparable to that of some critical housekeeping genes including *RIF1*, and *TET2* (**Figure 4-4**).

Total HML-2 transcript levels in Tera-1 cells were on a par with non-POU domain-containing, octamer-binding (*NONO*), a gene which encodes a DNA-and RNA-binding protein involved in transcriptional regulation and mRNA splicing (226). Total HML-2 levels in Hcc1954, HMLE-Ras, and HME were comparable to levels of replication timing regulatory factor 1 (*RIF1*), which encodes a protein involved in the DNA damage repair pathway (227), while total HML-2 levels in HMLE-Her2 were

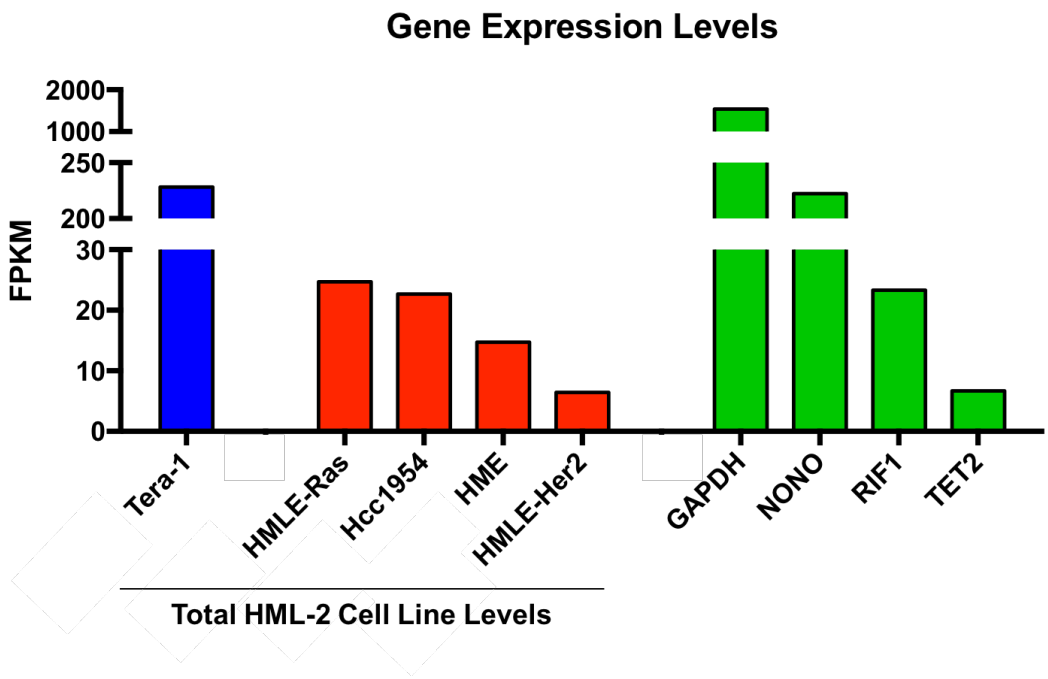


Figure 4-4. Total unique HML-2 transcript levels.
Bar graph showing the total unique HML-2 transcript levels in each cell line analyzed (red bars), including Tera-1 results from a study previously published by our lab (blue bar) (110). Comparisons are made against the transcript expression levels of housekeeping genes (GAPDH, NONO, RFI1, and TET2, green bars) in the Hcc1954 cell line.

comparable to tet methylcytosine dioxygenase 2 (*TET2*), which encodes an enzyme responsible for the conversion of 5-mC to 5-hmC during DNA demethylation (228). The level of glyceraldehyde 3-phosphate dehydrogenase (GAPDH) transcription, a very highly transcribed housekeeping gene, is shown for comparison (229).

The human reference genome (hg19) contains 91 annotated HML-2 full-length proviruses and 944 solo LTRs. This collection includes upwards of 30 human-specific proviruses, at least 11 of which are insertionally polymorphic (40, 109). Recent evidence shows that this list is not complete and that there are at least 36 additional polymorphic HML-2 sequences incorrectly annotated in hg19 (41). This list includes at least 4 full-

length proviruses that are listed as solo LTRs in hg19 as well as 5 full-length proviruses that are present as pre-integration sites (**Table 2-4**).

Although many non-reference HML-2 elements are extremely rare, their expression is important to detect as they may have major impacts on the physiology of the cell. These proviruses are the main candidates for aiding in a transformation event, since it is likely that proviruses with negative impacts on the host would have been selected against over the years and never reach fixation (3, 43, 99). To address this issue, we created an HML-2 reference genome that contains the sequences of all 96 currently known full-length proviruses, 976 solo LTRs, and 1 prototype SINE-R element, a type of retrotransposon that consists of the HML-2 *env* and 3' LTR sequences (198).

Two separate alignments (one to hg19 and one to our HML-2 reference genome) were conducted for the Hcc1954, HMLE-Ras, HMLE-Her2, and HME sequencing data and filtered for uniquely mapped reads only. All proviruses that contributed at least 1% of the total HML-2 transcript expression in a given cell line when aligning to either reference genome are shown in **Figure 4-5**. The reference genome comparisons show minimal differences between them, suggesting that there are no rare proviruses expressed in our HMEC transformation model. Of the seven human-specific proviruses with at least 1% abundance in these cell lines, none is insertionally polymorphic.

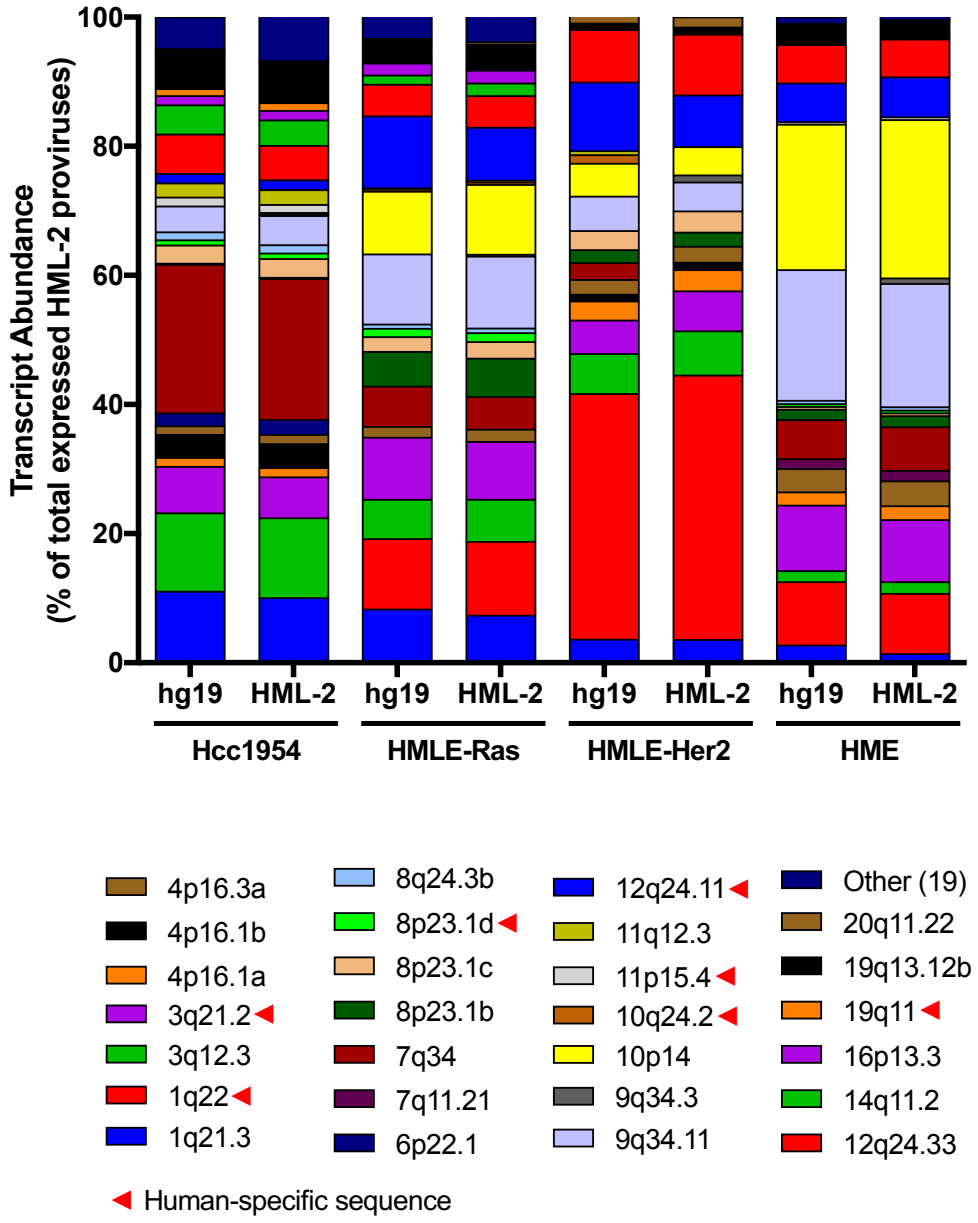


Figure 4-5. No expression of rare HML-2 proviruses.
Bar graph depicting the percent abundance of HML-2 mRNA when using two different reference genomes. Percent abundance is calculated as (provirus FPKM)/(total HML-2 FPKM) × 100. Cell lines sequenced as well as reference genomes used for alignment are listed under each bar. hg19, human reference genome; HML-2, HML-2 reference genome, which includes 9 additional polymorphic proviruses not present in the hg19 genome (**Table 2-4**). All proviruses contributing at least 1% of the total HML-2 expression are shown. Proviruses with <1% expression are grouped together as "Other". All proviruses known to be human-specific sequences are designated with a red triangle.

4.3 Expression is dominated by older proviruses producing antisense mRNAs

It is well established that tumorigenesis results in an increase in HML-2 expression in a large variety of human cancers (53, 105, 136, 154, 169, 230). However, the nature of this expression, in particular whether it is due to an increase in intensity of expression of the same proviruses or to an increase in diversity of expressed proviruses, is unclear. To address this question, we catalogued HML-2 expression in each sequenced cell line. We documented the transcript abundance level of each provirus as well as the orientation of transcription. All proviruses mentioned in this study are listed by chromosomal location, with aliases and genomic coordinates shown in **Table 4-2**.

Table 4-2. Transcribed HML-2 proviruses with aliases and genomic coordinates[a].

Provirus	Alternative Names	Chromosomal Location (hg19)
1q21.3		chr1:150,605,284-150,608,361
1q22	K102, K(C1b), K50a, ERVK-7	chr1:155,596,457-155,605,636
3q12.3	KII, ERVK-5	chr3:101,410,737-101,419,859
3q21.2	KI, ERVK-4	chr3:125,609,302-125,618,439
4p16.1b	K50c	chr4:9,659,580-9,669,174
4p16.3a		chr4:234,989-239,459
7q34	K(OLDAC004979), ERVK-15	chr7:141,451,918-141,455,938
8p23.1b	K27	chr8:8,054,700-8,064,221
8p23.1c		chr8:12,073,970-12,083,497
9q34.11	K31	chr9:131,612,515-131,619,736
10p14	K(C11a), K33, ERVK-16	chr10:6,866,141-6,875,603
12q24.11		chr12:111,007,843-111,009,325
12q24.33	K42	chr12:133,667,122-133,673,064
14q11.2	K(OLDAL136419), K71	chr14:24,480,600-24,484,985
19q13.12b	K(OLDAC012309), KOLD12309	chr19:37,597,549-37,607,066

[a]From Bhardwaj *et al.* (110), Subramanian *et al.* (40), and Gonzalez-Hernandez *et al.* (156). Only proviruses mentioned in the text or in figures are listed.

In total, we found only fifteen proviruses to be appreciably transcribed (FPKM >0.5) in any cell line sequenced (**Figure 4-6**). Surprisingly, only five of these were significantly sense-transcribed, with 75% of that transcription stemming from three loci (1q21.3, 1q22, and 3q12.3) (**Figure 4-7A**). In contrast, a diverse set of proviruses was expressed in antisense orientation (**Figure 4-7B**), suggesting that the majority of HML-2 transcription is due to *trans*-activating factors that result in a variety of non-coding transcripts.

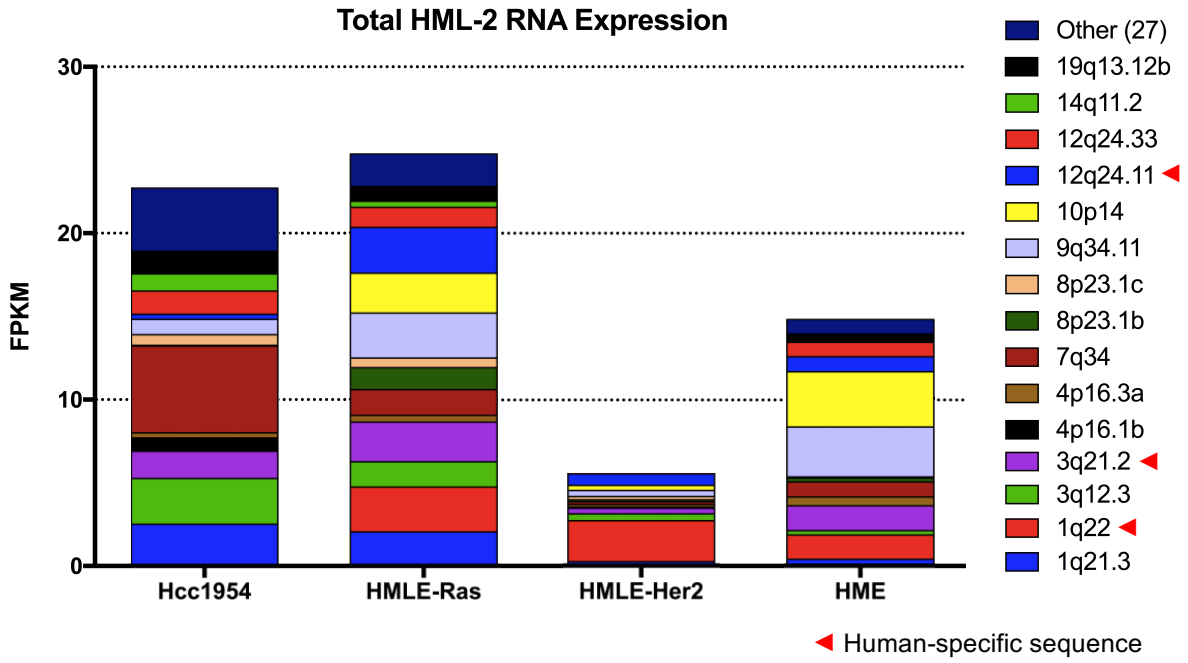


Figure 4-6. Multiple proviruses contribute to total HML-2 expression.
Bar graphs depicting the FPKM values of each significantly expressed provirus in the four cell lines sequenced. Significant expression is defined as proviruses with FPKM >0.5. Proviruses with FPKM <0.5 are grouped together as "Other". All proviruses that are known to be human-specific sequences, as described by Subramanian *et al.* (40), are designated with a red triangle.

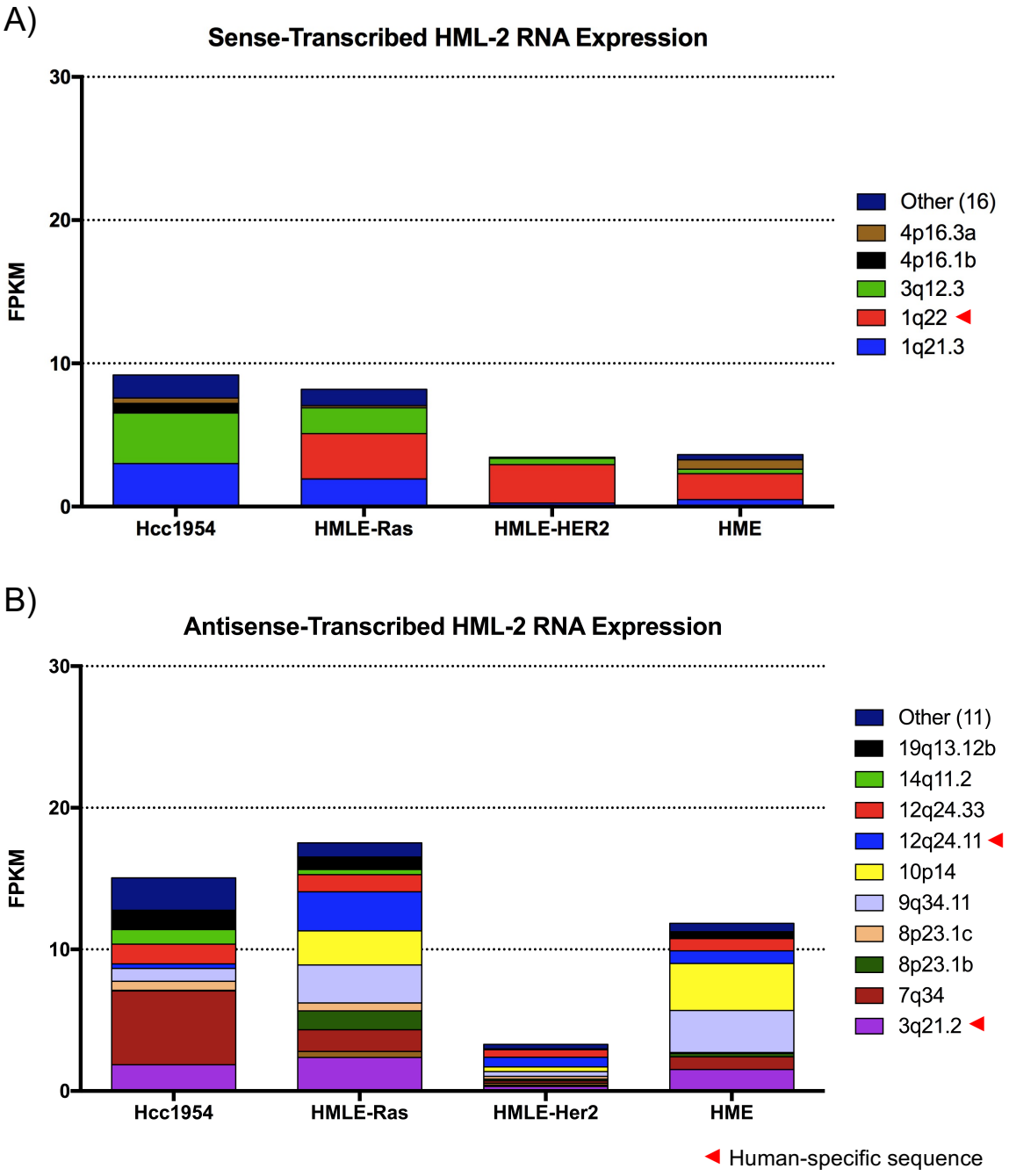


Figure 4-7. Proviruses transcribed in sense vs. antisense orientation.
Bar graphs depicting the FPKM values of each significantly expressed provirus in the cell lines sequenced that are transcribed in either (**A**) sense or (**B**) antisense orientation. Significant expression is defined as proviruses with FPKM >0.5. Proviruses with FPKM <0.5 are grouped together as "Other". All proviruses that are known to be human-specific sequences, as described by Subramanian *et al.* (40), are designated with a red triangle.

Overall, antisense transcription constituted the majority of expression in these cell lines: 72% of the total HML-2 expression in Hcc1954 cells, 73% in HMLE-Ras cells, 49% in HMLE-Her2 cells, and 82% in HME cells (**Figure 4-8**). Age analysis of each expressed provirus, as determined by the estimated time since integration, shows that all but one provirus (1q22) integrated at least five million years ago (**Figure 4-9**).

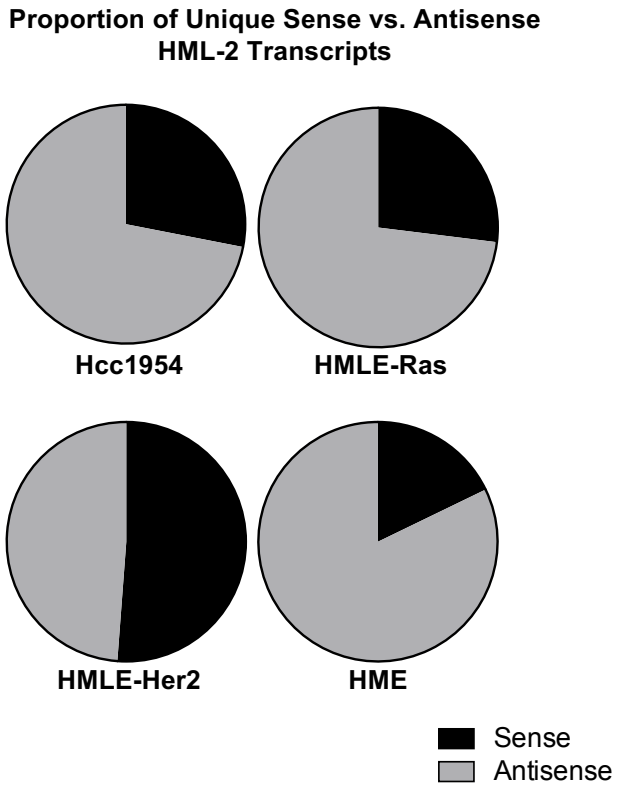


Figure 4-8. HML-2 expression is dominated by antisense transcription.
Pie charts comparing the percent of total HML-2 expression that is due to sense (black) vs. antisense (gray) transcription for each cell line sequenced. Sense was determined by strandedness of alignment.

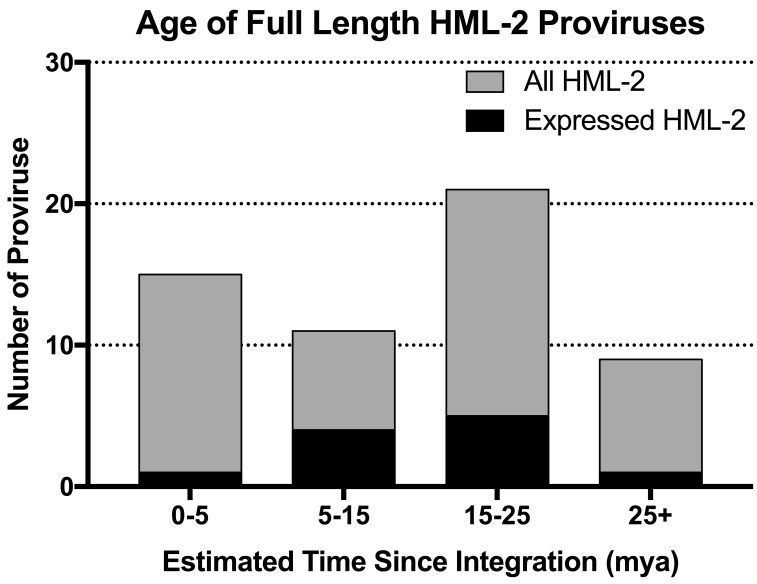


Figure 4-9. Age of expressed HML-2 proviruses.
Bar graph showing the age, given as time since integration in mya (million years ago), of full-length HML-2 proviruses. The total number of HML-2 proviruses within a given age range is shown in gray. Superimposed on top, in black, is the number of HML-2 proviruses with significant expression in this study. Age estimates are plotted as averages as determined by Subramanian *et al.* (40). Thirty-five proviruses have indeterminable ages and were excluded.

4.4 HML-2 transcription occurs via four mechanisms

We next chose to investigate the mechanism(s) behind HML-2 transcription and determine whether these differed with tumorigenicity. Through analysis of our alignments using the Integrative Genomics Viewer (200), we were able to detect HML-2 transcription patterns consistent with four different modes: initiation at the proviral LTR, read-through transcription, as part of a lncRNA, and intronic transcription (**Table 4-3, Figure 4-10**).

The majority of HML-2 antisense expression was associated with proviruses located in introns (**Figure 4-10A, right**). A schematic drawing, as well as a

Table 4-3. Characterization and identification of expression patterns for significantly expressed proviruses.

Antisense-Transcribed Proviruses

				Cell Line Expression			
Provirus	*LTRs*	*Mode of Transcription*	*Impacting Host Gene*	*Hcc 1954*	*HMLE-Ras*	*HMLE-Her2*	*HME*
§3q21.2	5'/Δ3'	Read-through	Rep. Element	+	+		+
7q34	Δ5'	Read-through	SSBP1	+	+		+
8p23.1b	5'/3'	Read-through	Rep. Element		+		
8p23.1c	5'/3'	Read-through	Rep. Element	+	+		
9q34.11	5'/3'	Intronic	CCBL1	+	+		+
10p14	5'/3'	Intronic	LINC00707		+		+
§12q24.11	5'	Intronic	PPTC7		+		+
12q24.33	5'/3'	Intronic	ZNF140	+	+	+	+
14q11.2	Δ5'/3'	Intronic	DHRS4L1	+			
19q13.12b	5'/3'	Intronic	ZNF420	+	+		

Sense-Transcribed Proviruses

				Cell Line Expression			
Provirus	*LTRs*	*Mode of Transcription*	*Impacting Host Gene*	*Hcc 1954*	*HMLE-Ras*	*HMLE-Her2*	*HME*
1q21.3	Δ5'/3'	Read-through	Rep. Element	+	+		
§1q22	5'/3'	lncRNA	BC041646		+	+	+
3q12.3	5'/3'	LTR-driven		+	+		
4p16.1b	5'/Δ3'	LTR-driven		+			
4p16.3a	5'	Intronic	ZNF876P				+

§Human-specific sequence; Δ, deletion; Rep. Element, repetitive element.

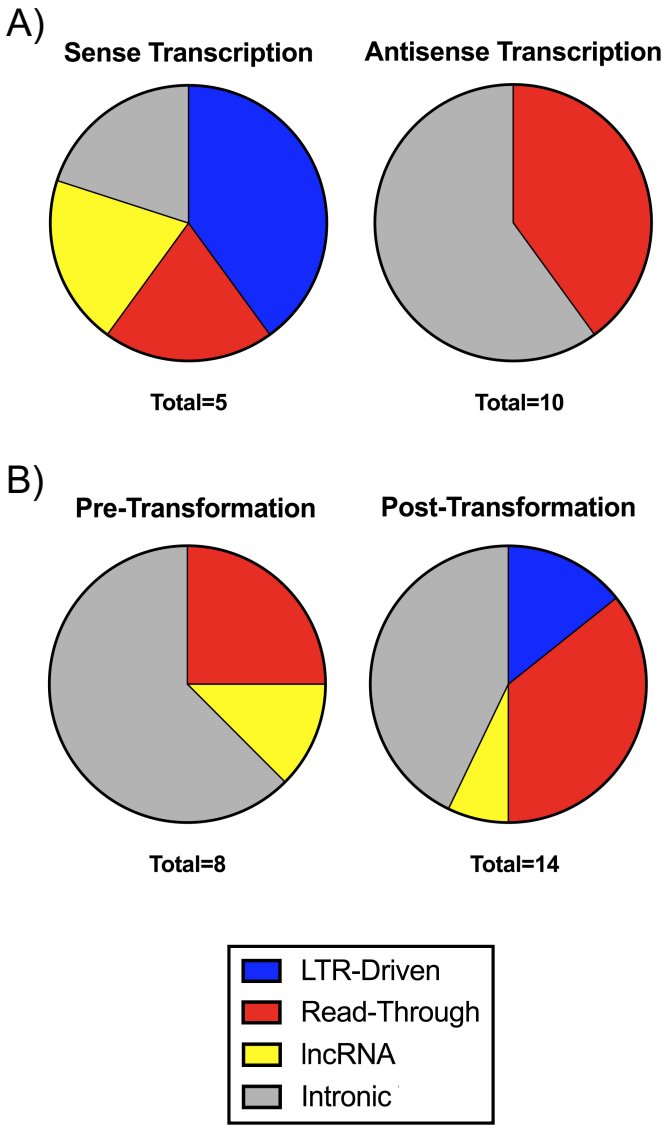


Figure 4-10. LTR-driven sense transcription is restricted to transformed cells. Pie charts showing the proportion of each mode of transcription identified through IGV to produce (**A**) sense (left) and antisense (right) transcripts as well as the proportion of each mode of transcription used in (**B**) cell lines pre-transformation (left, includes HME cell line) and post-transformation (right, includes HMLE-Ras, HMLE-Her2, and Hcc1954).

representative IGV visualization, of this mode of transcription is seen in **Figure 4-11A**, whereby proviral alignments are evident within an intronic region of a host gene. This correlation suggests that these proviruses are not self-transcribed, but rather that their expression is seen as a consequence of being preserved within an incompletely removed intron. All other antisense HML-2 transcription was due to read-through (**Figure 4-10A, right**), from being situated downstream of a transcribed host gene or repetitive element. A representative image of this mechanism is shown in **Figure 4-11B**, whereby transcription is seen continuing past the last host gene exon and into the proviral sequence.

Most interestingly, three of the sense-transcribed proviruses appear to have functional LTRs. 1q22 is expressed as part on an annotated lncRNA of unknown function (BC041646), known to be highly expressed in breast epithelium as well as white blood cells, placenta, lung, and lymph node (112, 184, 231). This lncRNA originates from an upstream (TCC)n simple repeat element and terminates using the polyadenylation signal found in the 1q22 3' LTR (**Figure 4-12A**). Two proviruses, 3q12.3 and 4p16.1b, appear to have LTR-driven transcription, whereby reads are seen originating from the transcription start site. A schematic of LTR-driven transcription is shown in **Figure 4-12B**. Interestingly, LTR-driven transcription was the only mechanism of expression related to tumorigenicity, as it was not detected in non-tumorigenic cells (**Figure 4-10B**). These results suggest that the transcriptional milieu of a tumorigenic cell is critical for LTR promoter activity, corroborating the previous HML-2 reporter construct assay results (**Figure 3-3**).

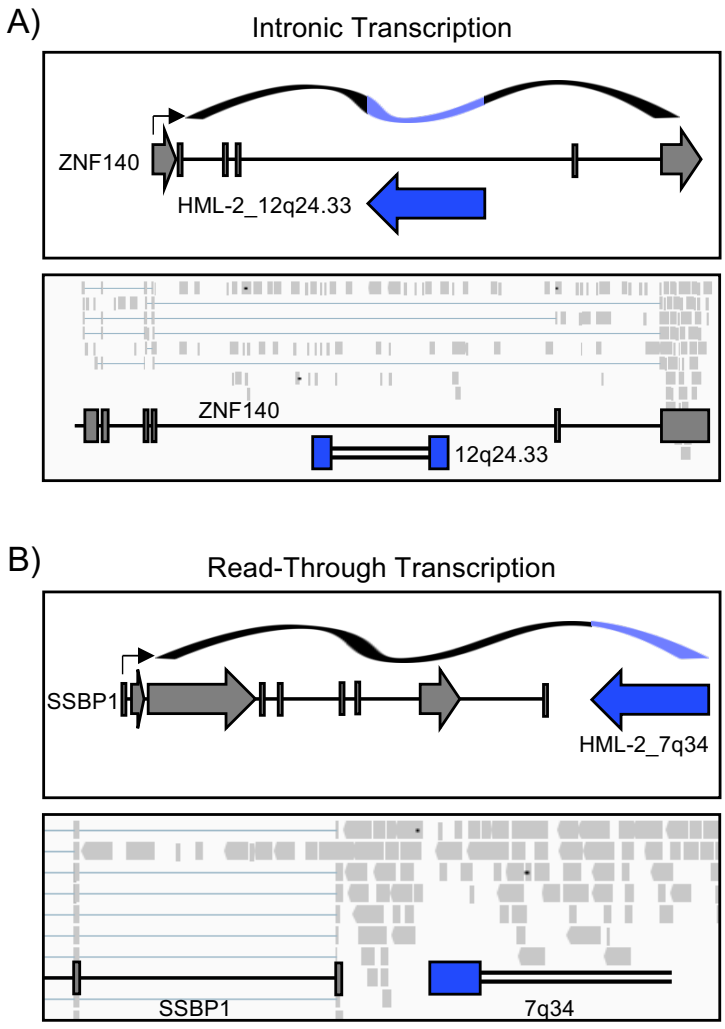


Figure 4-11. Examples of intronic and read-through proviral transcription. Images depicting examples of two of the four mechanisms of HML-2 transcription as identified in our data set: (**A**) intronic transcription and (**B**) read-through transcription. Schematic drawings of each mode of transcription are shown on the top. Dark gray boxes depict host genes and arrows correspond to direction of transcription. Blue boxes depict proviral sequences and arrows depict direction of sense transcription. Black arrows designate transcriptional start sites. Top ribbon depicts RNA with black sections signifying the sequences of host genes and blue sections signifying proviral sequences. Representative IGV visualizations of each mode of transcription are shown on the bottom. Light gray boxes show aligned sequencing reads, which range from 60-592 bp in length. Splicing is shown by light blue lines connecting reads. Location of host and proviral genes are superimposed on top of alignments. Blue boxes depict LTR sequences and double black lines depict internal proviral sequences.

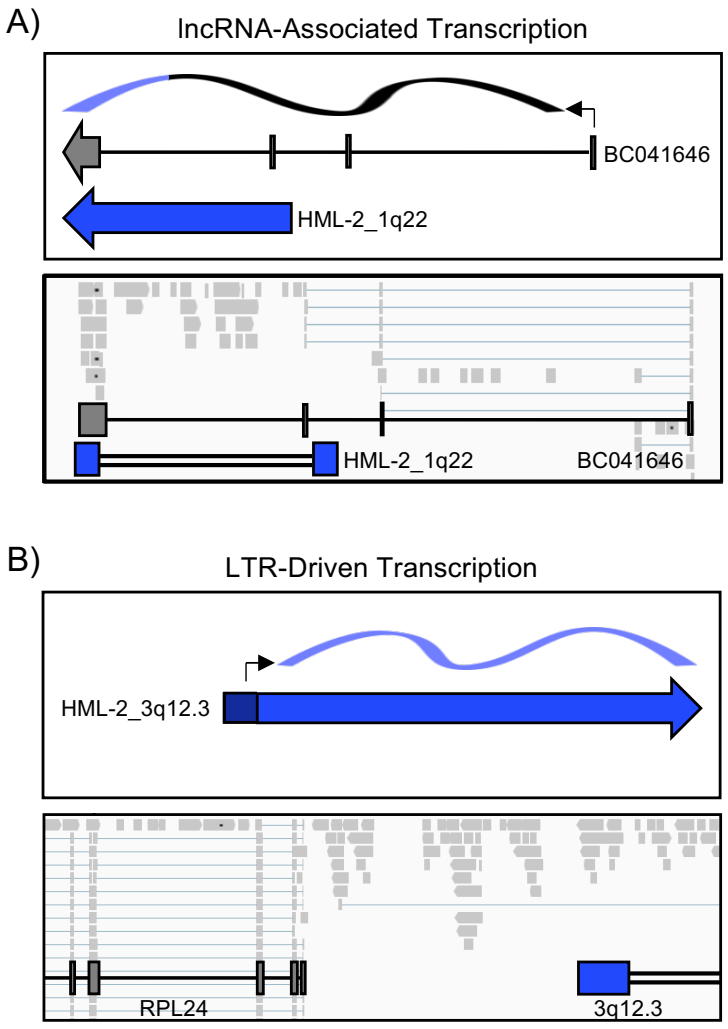


Figure 4-12. Examples of lncRNA-associated and LTR-driven proviral transcription.
Images depicting examples of two of the four mechanisms of HML-2 transcription as identified in our data set: (**A**) lncRNA-associated transcription and (**B**) LTR-driven transcription. Schematic drawings of each mode of transcription are shown on the top. Dark gray boxes depict host genes and arrows correspond to direction of transcription. Blue boxes depict proviral sequences and arrows depict direction of sense transcription. Black arrows designate transcriptional start sites. Top ribbon depicts RNA with black sections signifying the sequences of host genes and blue sections signifying proviral sequences. Representative IGV visualizations of each mode of transcription are shown on the bottom. Light gray boxes show aligned sequencing reads, which range from 60-592 bp in length. Splicing is shown by light blue lines connecting reads. Location of host and proviral genes are superimposed on top of alignments. Blue boxes depict LTR sequences and double black lines depict internal proviral sequences.

4.5 No detectable effect of functional 5' LTRs on host gene transcription

Confirmation of 3q12.3 and 4p16.1b 5' LTR promoter activity was achieved through a dual-luciferase assay (207). As shown in **Figure 4-13**, the 3q12.3 5' LTR exhibited high promoter activity in all cell lines sequenced except for HME. This activity correlated with the transcript expression seen in Hcc1954 and HMLE-Ras cells. Interestingly, the LTR had high promoter activity in HMLE-Her2 cells, despite the low level of 3q12.3 transcripts produced by that cell line (**Figure 4-13A**). The 4p16.1b LTR showed a detectable, but low level of promoter activity in Hcc1954, which correlated with transcript expression seen in only those cells (**Figure 4-13B**), implying the absence of necessary *trans*-acting factors in the other three cell lines sequenced.

There is evidence that retroviral LTRs are capable of influencing host gene transcription up to 100 kb away (3, 4). To investigate whether the active 3q12.3 or 4p16.1b LTRs were influencing host gene transcription, we compared the gene expression levels of all host genes, as annotated by RefSeq in the UCSC Genome Browser, within 100 kb in either direction (112). We found no known genes in the vicinity of 4p16.1b, ruling out its possibility of affecting host gene transcription. There are seven genes (four upstream and three downstream) within 100 kb of the 3q12.3 provirus. However, we found no correlation between provirus expression and host gene expression in any cell line sequenced (**Figure 4-14**). From this result, we concluded that HML-2 LTR activity had no detectable effect on host gene transcription in these samples.

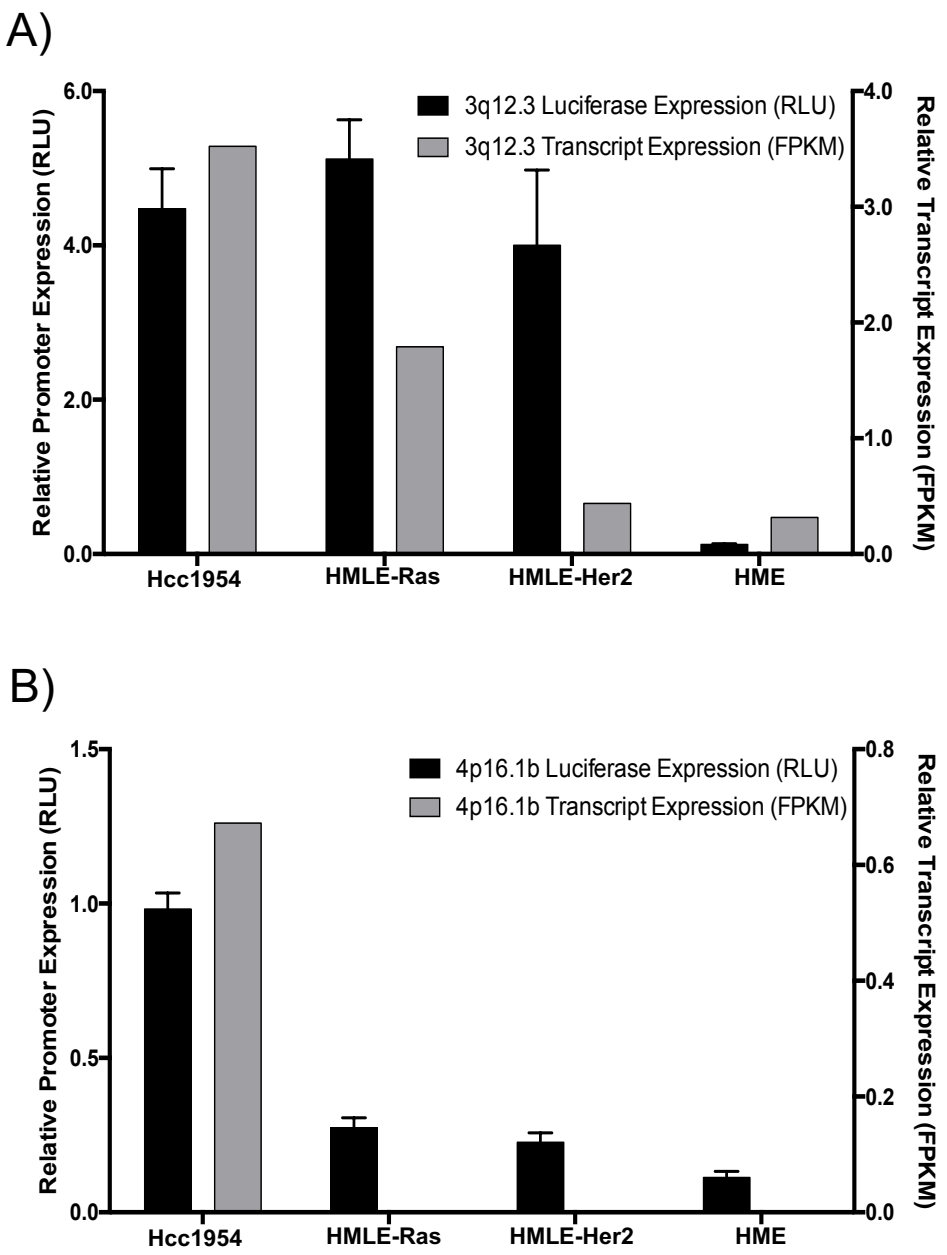


Figure 4-13. Validation of active 5' LTRs.
Confirmation of 5' LTR activity of the two proviruses shown to have LTR-driven sense transcription in our data set. Promoter activity is given as relative light units (RLU, black) of the LTR-driven luciferase activity normalized to a co-transfected control of *Renilla* luciferase activity driven by an SV40 promoter. Relative transcript expression is given as FPKM (gray). FPKM values are from stranded Cuffdiff analyses and are normalized across all cell lines. Results for 3q12.3 are shown in (**A**) and 4p16.1b are shown in (**B**).

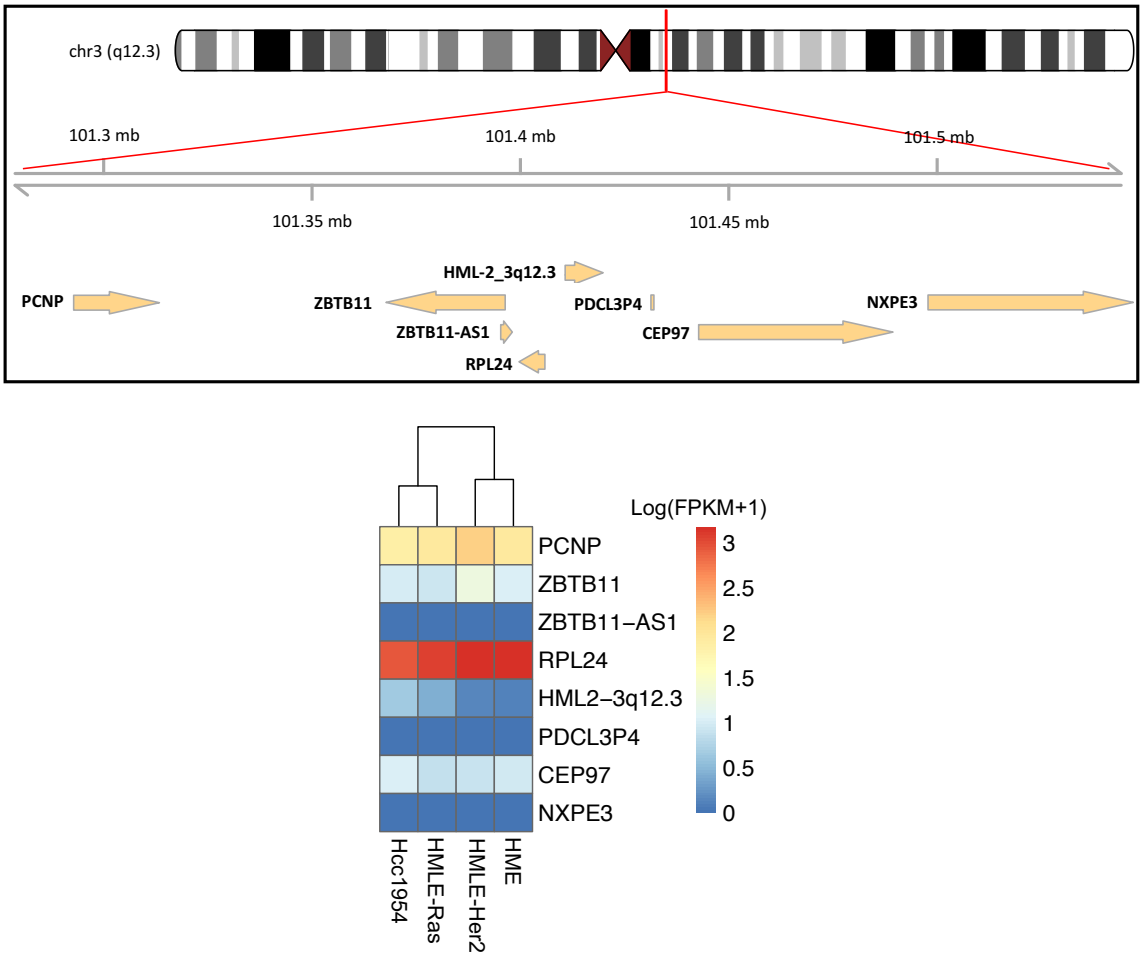


Figure 4-14. 3q12.3 5' LTR has no discernible impact on host gene transcription. Top, schematic showing the 3q12.3 locus including all genes found within 100 kb of either side of the provirus as annotated by RefSeq in the UCSC Genome Browser. Bottom, heatmap showing the expression levels of each gene within each cell line sequenced. Values are given as log(FPKM+1) and are normalized through Cuffdiff analysis. Clustering is based on Euclidean distance.

Chapter 5: The relationship between tumorigenic cellular environment and LTR sequence variation

5.1 HML-2 5' LTR activity in Tera-1 cells

The association between LTR sequence and cell line-specific expression suggests that certain sequence-specific elements play a large role in determining differential promoter activity (**Figure 3-6**). Recent evidence suggests that HML-2 LTRs act in an Inr- and TATA-independent manner, and instead rely upon specific *cis*-acting regulatory sequences found in the U3 region of the 5' LTR to drive transcription (162, 164). Furthermore, *in vitro* studies exposing cell lines to demethylating agents show that hypomethylation alone is not sufficient to induce LTR promoter activation, suggesting that the proper transcriptional milieu of a cell is critical for transcription initiation (151, 162, 172).

We first chose to explore the importance of the U3 region by investigating 5' LTR activity in Tera-1 cells. The transcript abundance levels of HML-2 proviruses was determined by Neeru Bhardwaj (110). She found that 50% of the HML-2 transcription in Tera-1 cells originated from two proviral loci: 22q11.21 and 22q11.23 (**Figure 5-1A**). Curiously, in our attempt to validate the 5' LTR activity of these proviruses using a reporter construct assay, we found that transcript abundance and promoter activity levels did not correlate. Although the 22q11.21 5' LTR exhibited high promoter activity, accounting for its high transcript abundance, the 22q11.23 5' LTR exhibited no promoter activity in this cell line (**Figure 5-1B**). The 22q11.21 provirus is human-specific, belongs to the LTR-Hs subgroup, and is relatively young with an estimated age of 1.84-3.34 million years old (40, 112, 232). In contrast, the 22q11.23 provirus is estimated to be

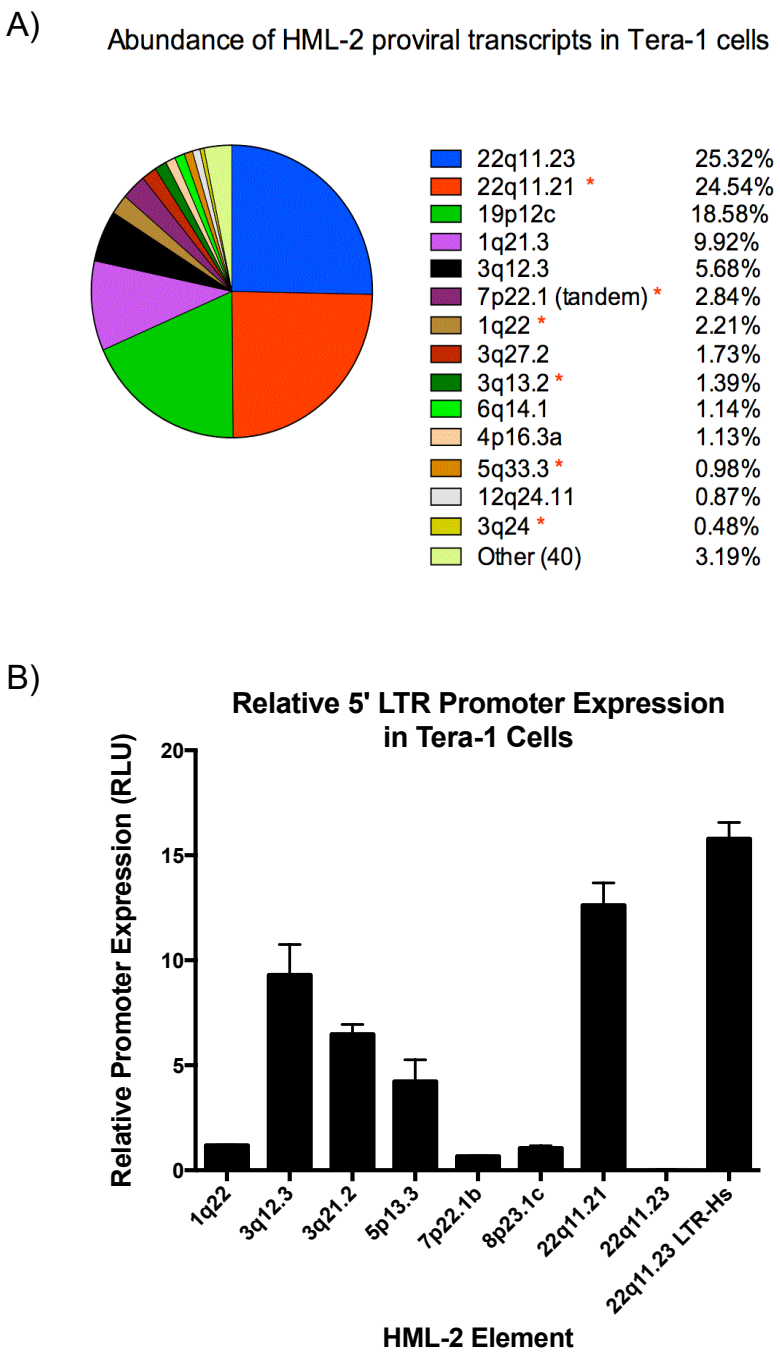


Figure 5-1. HML-2 transcript abundance and promoter activity levels in Tera-1 cells.
(**A**) Transcript abundance levels in Tera-1 cells, as determined by RNA-Seq analysis accomplished by Neeru Bhardwaj. Abundance levels reflect data from unique, stranded alignments to hg19. Proviruses with a red asterisk are human-specific. Reproduced with permission of Multidisciplinary Digital Publishing Institute (110). (**B**) Relative promoter activity of select HML-2 5' LTRs in Tera-1 cells. Promoter activity is determined as relative light units (RLU) normalized against the internal control *Renilla* expression. All experiments were conducted in triplicate and data display the mean ± standard deviation.

much older. Its 5' LTR belongs to the most ancestral LTR5B subgroup and the provirus is believed to be 21.64-39.18 million years old, having first integrated in genome of our last common ancestor with gorillas (40, 233). Although the old age of 22q11.23 may explain the lack of retention of promoter activity, as it has most likely acquired a large number of mutations since its time of integration, it does not explain the extremely high transcript abundance level seen in Tera-1 cells (110).

IGV analysis of the RNA-Seq alignments showed that the 22q11.23 transcripts were not driven off of its native 5' LTR. Instead, reads were originating from an upstream HML-2 element, located 490 bp away from the provirus. This element, which we deemed 22q11.23 LTR-Hs, is not quite a solo LTR. Instead, it consists of a full 5' LTR-Hs sequence followed by 227 bp of downstream *gag* sequence (**Figure 5-2**). The alternative names and chromosomal locations of the elements found at the 22q11.23 locus are listed in **Table 5-1.** Being of the LTR-Hs subgroup, the 22q11.23 LTR-Hs

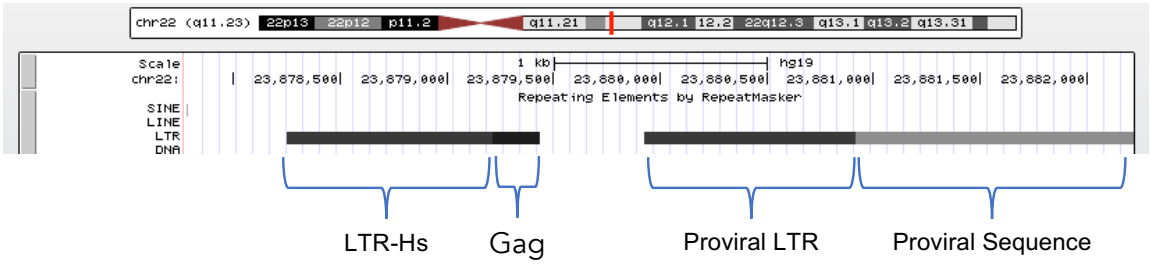


Figure 5-2. Schematic of the 22q11.23 locus.
Schematic of the 22q11.23 locus, as shown by the RepeatMasker track on the UCSC Genome Browser (112). Image shows the upstream 22q11.23 LTR-Hs element, consisting of a full length 5' LTR sequence, followed by a 227 bp truncated *gag* sequence. The 22q11.23 full-length provirus is located 490 bp downstream of the LTR-Hs element. Only the proviral 5' LTR and a portion of downstream proviral sequence are shown for the 22q11.23 provirus.

Table 5-1. Alternative names and genomic coordinates of HML-2 elements found at the 22q11.23 locus[a].

Element	Alternative Names	Chromosomal Location (hg19 build)
22q11.23 Provirus	K(OLDAP000345), KOLD345	chr22: 23,879,927-23,889,087
22q11.23 LTR-Hs		chr22: 23,878,249-23,879,376

[a]From Subramanian *et al.* (40) and Bhardwaj *et al.* (110).

element is presumed to be quite young in relation to other HML-2 integrations (40). Reporter construct analysis indicated that this LTR had very high promoter activity in Tera-1 cells, comparable to those exhibited by the 22q11.21 5' LTR (**Figure 5-1B**).

Due to its high promoter activity, we chose to use the 22q11.23 LTR-Hs element to investigate the importance of the U3 region in driving transcription. Genomic analysis of this LTR showed that it still possessed a large number of core promoter elements, including multiple enhancer sequences, a TATA box, and an Inr element. IGV analysis showed that the main transcription start site (TSS) was located at nt position 826, but that there were at least four alternative TSS (**Figure 5-3A, top**). As research suggests that HML-2 LTRs are TATA- and Inr-independent, we sought to investigate whether the removal of these sites would impact LTR promoter activity. To assess this, we created seven truncated versions of this LTR by removing varying amounts from the 3' end of the sequence (**Figure 5-3A, bottom**). These truncated versions were inserted into promoter-less luciferase vectors and assessed through the same dual-luciferase assay as reported previously (see Materials and Methods).

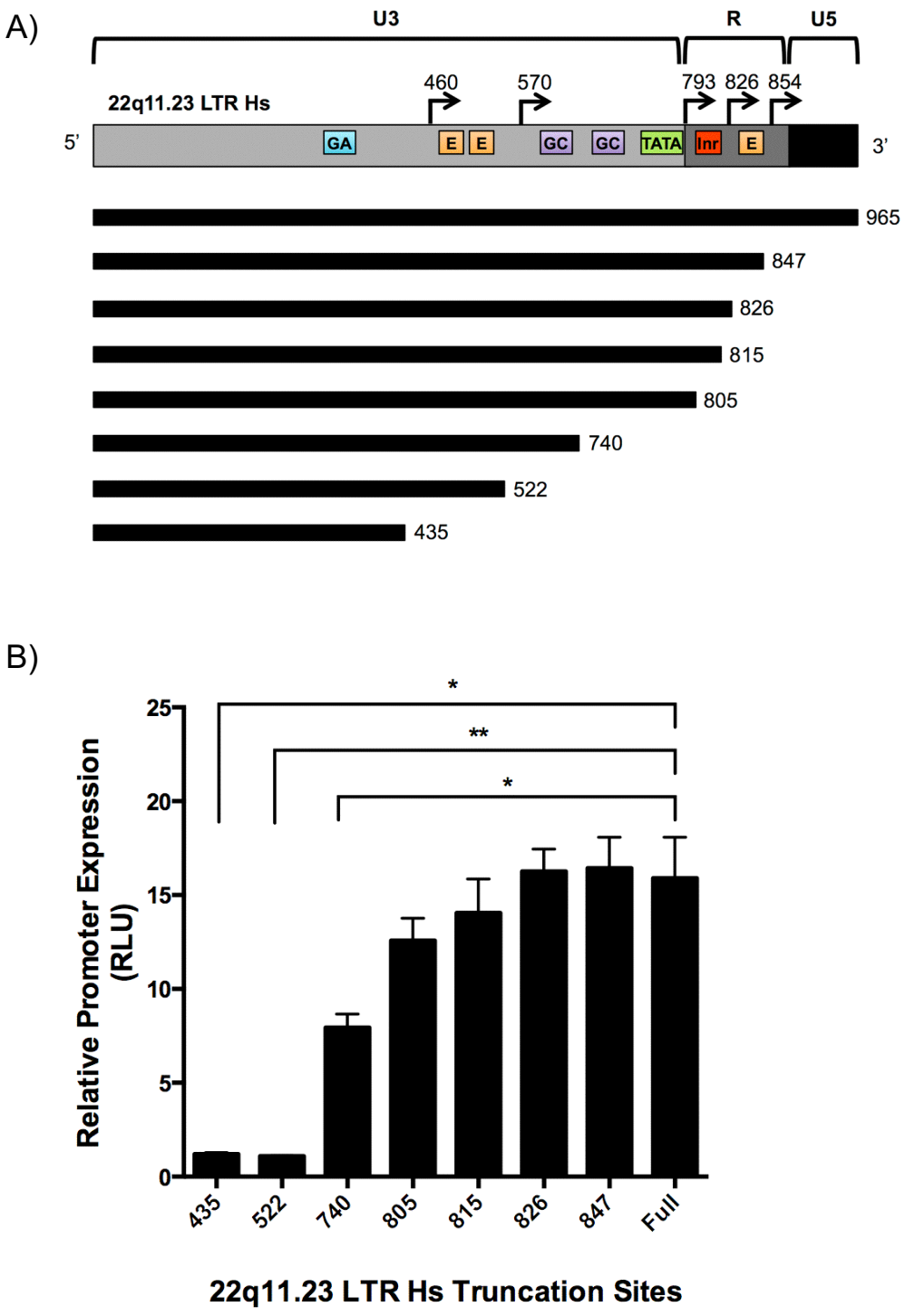


Figure 5-3. LTR promoter activity is not dependent upon R or U5 sequences. Full figure legend is continued on the next page.

Figure 5-3. LTR promoter activity is not dependent upon R or U5 sequences. (**A**) Top, schematic of the 22q11.23 LTR-Hs 5' LTR with U3, R, and U5 regions labeled. Predicted TSS, as determined through RNA-Seq analysis, are shown with arrows and numbered by nucleotide position. Colored boxes indicate previously described core promoter element motifs (162, 204, 234). GA, GA rich motif (nt 379–386, sequence GGGAAGGG); E, enhancer box (nt 465–476, sequence TTGCAGTTGAGA; nt 485–496, sequence AGGCATCTGTCT; nt 832–843, sequence CTCCATATGCTG); GC, GC rich motif nt 759–763, (sequence CCCCC; nt 602–606, sequence GGCGG); TATA, TATA box (nt 790–797, sequence AATAAATA); Inr, initiator element (nt 807–812, sequence CTCAGA). Bottom, black lines indicate the regions included in each truncated LTR construct and numbers to the right of each line indicate the nucleotide position at which the LTR was truncated. Figure is not drawn to scale. (**B**) Relative promoter activity of the 22q11.23 LTR-Hs full-length and truncated constructs in Tera-1 cells (Kruskal-Wallis, *$p<0.05$, **$p<0.01$). Promoter activity is determined as relative light units (RLU) normalized against the internal control *Renilla* expression. All luciferase experiments were conducted in triplicate and data display the mean ± standard deviation. Reproduced with permission of Multidisciplinary Digital Publishing Institute (110).

Results indicated that truncating the LTR down to 805 nt, which resulted in removal of the entire U5 region as well as the majority of the R region including the main TSS and Inr element, did not significantly decrease promoter activity. A 2-fold significant reduction was seen after truncating down to 740 nt, removing the entire R-U5 region as well as ~50 bp of the U3 region including the TATA box sequence. However, complete reduction of promoter activity was not observed until the LTR was truncated down to 522 bp, removing another 200 bp of the U3 region as well as all but one alternative TSS (**Figure 5-3B**). These results indicated that 5' LTR activity is not dependent upon core promoter elements located in the R or U5 regions of the LTR, including the Inr element. Although it is not possible from these results to tell if removal of the TATA box resulted in the reduction of promoter activity, LTR functionality does appear to be reliant upon the sequences located in the 3' half of the U3 region. We used these preliminary results to investigate which specific sequences in the U3 region of the

HML-2 LTR are important for driving promoter expression in our HMEC transformation model.

5.2 Identification of binding sites critical for HML-2 promoter activity

Increased HML-2 activity is widely documented to occur during tumorigenesis (53, 105, 110, 136). Although this upregulation is corroborated by our RNA-Seq results, we also showed that HML-2 transcription is present in non-diseased cells. However, this latter activity was found to be dependent upon indirect methods of transcription as all provirus expression in HME cells was driven by an upstream promoter not associated with the LTR (**Figure 4-10B, left**). In contrast, LTR-driven transcription was only found to occur post-transformation (**Figure 4-10B, right**), suggesting that the malignant shift in transcriptional environment of the cell was critical for LTR promoter activation.

We chose to investigate this phenomenon further and to elucidate which LTR sequences were required for neoplastic promoter expression. We focused on three cell lines (HME, HMLE-Her2, and HMLE-Ras) that are derived from primary HMECs (**Figure 4-1**) and differ by oncogene overexpression (**Figure 4-2**). RNA-Seq results demonstrated significant differences in the global gene expression of each cell line (**Figure 4-3**). This model provided the opportunity to investigate how specific differences in the transcriptional environment of a cell can affect LTR expression.

We first assessed LTR promoter activity in these cell lines by using the same reporter construct assay as described in Chapter 3, and repeating the experiment with the same nine 5' LTRs as listed in **Table 3-2**. We detected increased 5' LTR activity from the 3q12.3 and 11p15.4 proviruses in the HMLE-Ras cell line as well as increased

activity from the 3q12.3 5' LTR in the HMLE-Her2 cell line. The significance of this expression was determined as compared to HME cells (**Figure 5-4**). In an effort to explain this pattern of expression, we sought to identify transcription factor binding sites that were unique to each LTR and therefore may be responsible for the selective activation seen of one 5' LTR over another in the HMLE cell lines. Using MatInspector, a transcription factor binding site prediction software by Genomatix (201), we found a total of 63 unique sites among the nine LTRs in this study. Of those, 13 were unique to the 3q12.3 5' LTR and 20 were unique to the 11p15.4 5' LTR (**Table 5-2**).

The MatInspector software was additionally used to create a list of transcription factors known to bind to these unique sites, as determined by review of current literature. The transcript abundance levels, given as FPKM, of these transcription factors in the Hcc1954, HMLE-Ras, HMLE-Her2, and HME cell lines was determined using data from our RNA-Seq experiment described in the previous chapter. These FPKM values were compared to assess if the expression of any of those transcription factors of interest were upregulated in the tumorigenic cell lines as compared to the non-transformed control, as increased expression could account for the increase in promoter activity of the LTRs containing those binding sites. Overall, we saw a significant increase in expression of transcription factors known to bind to the HOX-PBX and RFX3 sites on the 3q12.3 5' LTR (**Figure 5-5, left**) as well as a significant increase in those known to bind to the ATF and RORA sites on the 11p15.4 5' LTR (**Figure 5-5, right**), suggesting that these sites may be responsible for LTR activation in HMLE-Ras and HMLE-Her2 cells. These binding sites are located in similar regions of each LTR, with one being situated just

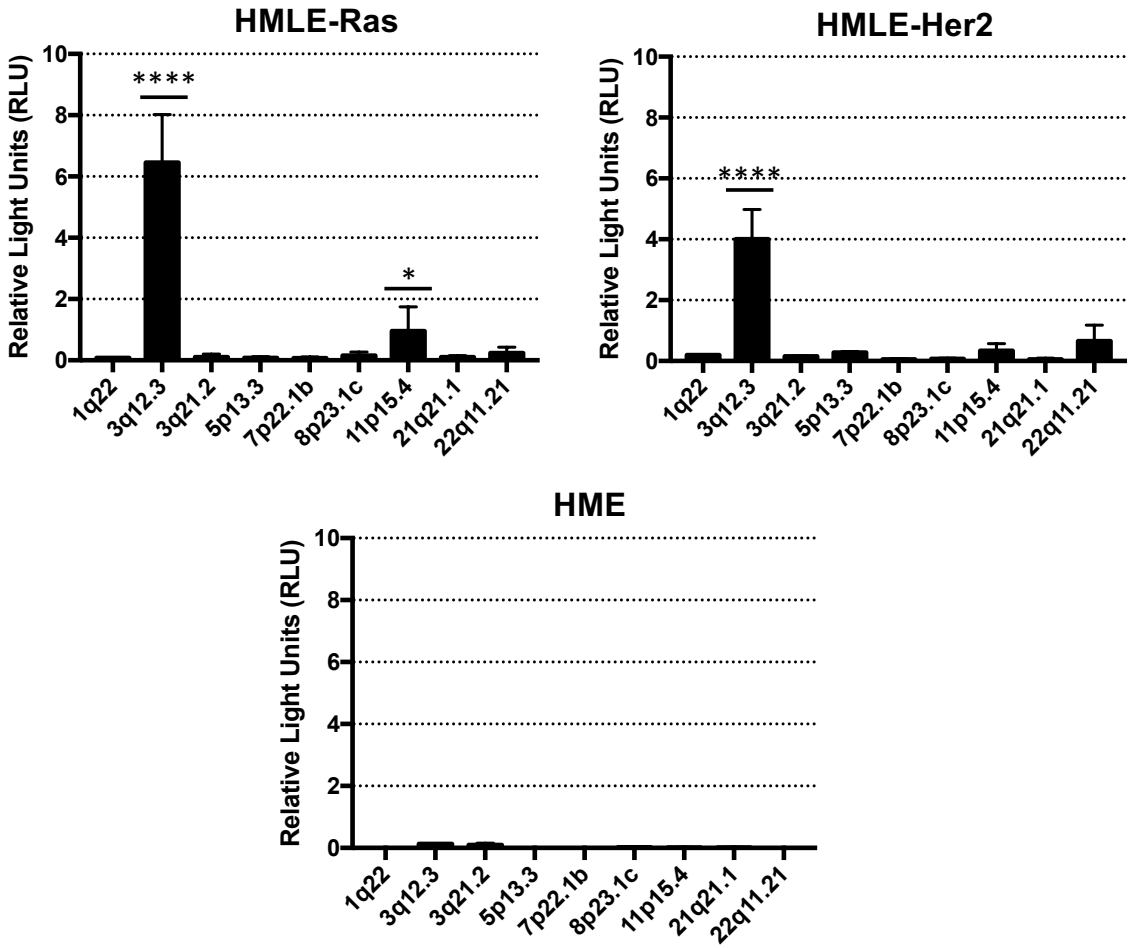


Figure 5-4. Promoter activity in HMLE and HME cells.
Relative 5' LTR promoter activity in the tumorigenic HMLE-Ras and HMLE-Her2 cell lines (top) as well as the immortalized, non-transformed HME cell line (bottom). Promoter activity is determined as relative light units (RLU) normalized against the internal control *Renilla* expression. Statistical significance was generated by ANOVA with Bonferroni's multiple comparisons test and is based on comparisons to HME expression (*$p<0.05$, ****$p<0.0001$). All luciferase experiments were conducted in triplicate and data display the mean ± standard deviation.

Table 5-2. Unique transcription factor binding sites in HML-2 5' LTRs of interest[a].

Provirus	Unique Binding Site(s)
1q22	NBRE‡
3q12.3	CDE, E2F, HOX-PBX, MRG1‡, PROX1, RFX3†, SF1†, SOX10, TAL1-E2A, TBX20, TEF-HLF‡, TGIF†, TR2‡
3q21.2	GLI1†, IK3, NFY‡, NKX29†, SIX2‡, STAT5
5p13.3	CARF‡, MYBL1‡
7p22.1b	EKLF‡, GAGA‡, GLI3‡
8p23.1c	AML1‡, BHLHB2‡, DMRT7‡, HMGA‡, HOX1-3‡, MAFF‡, MEF2‡, NRF1‡, PAX1‡, SOX17‡, STAT5A‡
11p15.4	AHRARNT‡, ATF, ATF6, CETS1P54, EN2‡, ETS1, FOXP2†, FREAC2‡, HDBP1-2, HHEX, HIVEP1†, IRF5, MIF1†, NRL, PEGASUS, RORA, SGN1, STAT3, TEF†, TGIF2LX
21q21.1	CHOP†, NFKAPPAB50†, USF†, ZNF300†
22q11.21	GRHL1‡, MASH1†, TAL1BETAHEB†

[a]Only unique sites for each 5' LTR, as compared to the other eight, are shown
† present only in other solo LTR(s)
‡ present only in other full-length provirus(es) and solo LTR(s)

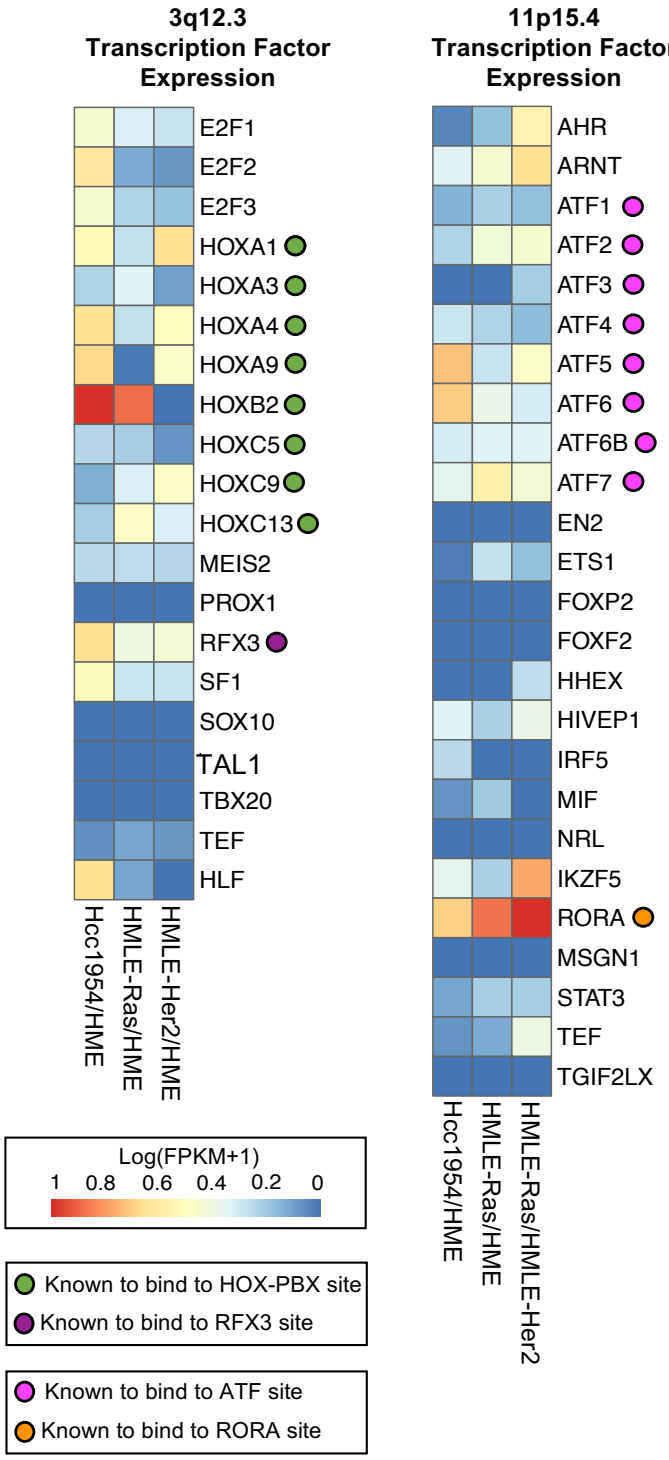


Figure 5-5. Identification of transcription factor binding sites critical for HML-2 promoter activity in HMLE cells.
Heatmaps showing the relative transcript abundance levels, in FPKM, of transcription factors known to bind to the unique sites on the 5' LTR of 3q12.3 (left) and 11p15.4 (right). Values are shown as the log (FPKM+1). Transcription factors known to bind to each site are designated by circles as color coded in the legend.

upstream of the U3-R boundary and the other being situated closer to the middle of the U3 region (**Figure 5-6**). The locations of these binding sites are consistent with the results from our LTR truncation experiment, which suggested that sequences found in the 3' half of the U3 region are critical for driving proviral transcription (**Figure 5-3**).

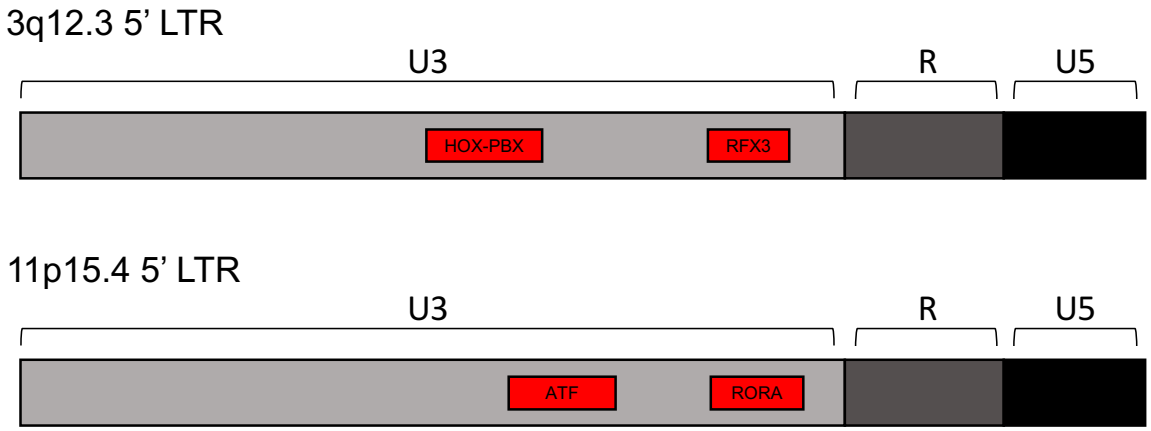


Figure 5-6. Location of critical binding sites on each 5' LTR.
Approximate locations of the HOX-PBX and RFX3 transcription factor binding sites on the 3q12.3 5' LTR (top) and the ATF and RORA transcription factor binding sites on the 11p15.4 5' LTR (bottom). U3, R and U5 regions for each LTR are shown. Figure is not drawn to scale.

5.3 Removal of critical binding sites decreases HML-2 promoter activity

The functionality of these putative LTR enhancer sites was assessed by mutating each one individually. A multiple sequence alignment was performed using the sequences of all nine 5' LTRs used in the reporter construct assay. From this alignment, we created a consensus sequence for each critical binding site, which we deemed to be the "non-active" version of each site (**Figure 5-7**). The 3q12.3 HOX-PBX binding site differed

from the consensus sequence by five nucleotides: a four bp insertion found in the middle of the site as well as one point mutation in the last nucleotide (**Figure 5-7, top**). Mutation of this site back to the consensus version significantly decreased LTR promoter activity in both neoplastic cell lines, with activity decreasing by 2-fold in HMLE-Ras cells (**Figure 5-8A**) and by 7-fold in HMLE-Her2 cells (**Figure 5-8B**).

The 3q12.3 RFX3 site differed from the consensus sequence by only one nucleotide (**Figure 5-7, bottom**), and yet removal of this site decreased LTR activity by 5-fold in both HMLE-Ras cells (**Figure 5-8A**) and HMLE-Her2 cells (**Figure 5-8B**). Activity was decreased to levels comparable to that of 1q22, a proviral LTR with no significant promoter activity in these cell lines (**Figure 5-8A, B**). By comparison, mutating these sites did not significantly decrease LTR promoter activity in Hcc1954 cells (**Figure 5-8C**). This tumorigenic cell line showed elevated expression of transcription factors known to bind to five unique 3q12.3 sites, including E2F, HOX-PBX, RFX3, SF1, and TEF-HLF (**Figure 5-5, left**). These results suggest that the other active transcription factors present in Hcc1954 cells were able to compensate for promoter activity when one site was removed.

Similar results were seen with the 11p15.4 5' LTR. The consensus sequence differed from the ATF binding site by seven nucleotides: five of which resulted in transition point mutations (substitution of a purine with another purine, or substitution of a pyrimidine with another pyrimidine) and two of which resulted in transversion point mutations (substitution of a pyrimidine with a purine, or vice versa) (**Figure 5-9, top**). Removal of this site decreased promoter activity by 6-fold in HMLE-Ras cells (**Figure 5-**

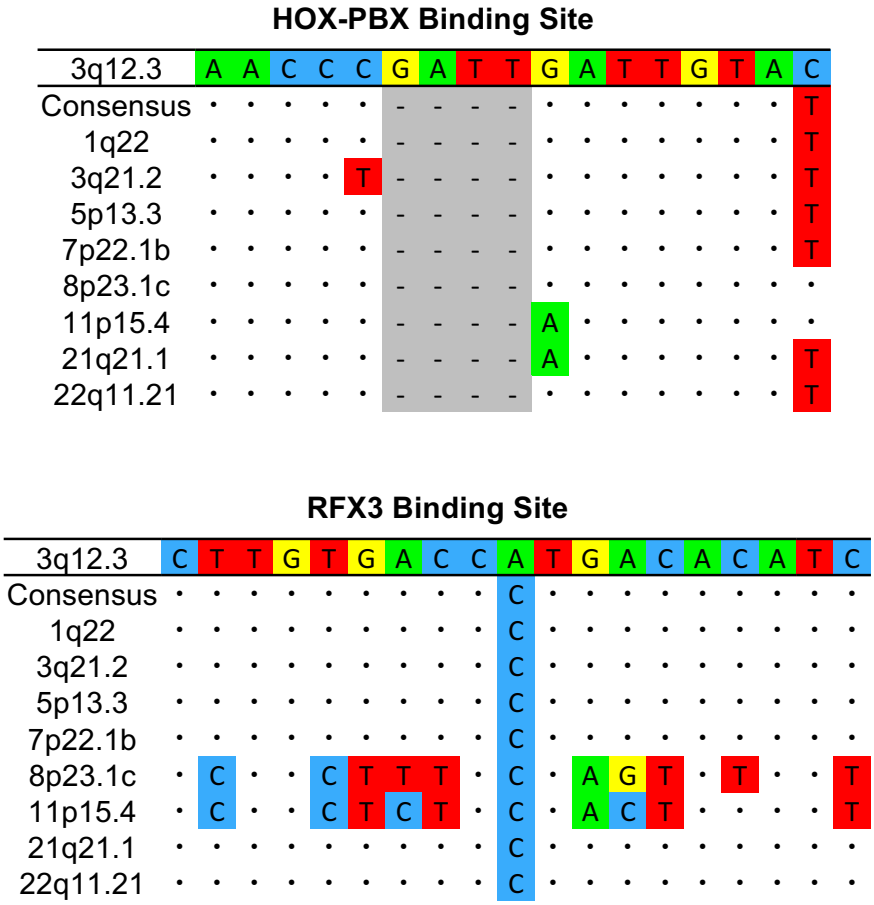


Figure 5-7. Multiple sequence alignments of the HOX-PBX and RFX3 binding sites on the 3q12.3 5' LTR.
Multiple sequence alignment of the HOX-PBX (top) and RFX3 (bottom) binding regions on the nine 5' LTRs of interest in this study as well as a consensus sequence of the site. Sequences are compared against the 3q12.3 5' and 3' LTR sites, with dots used for shared identity and dashes used to indicate indels.

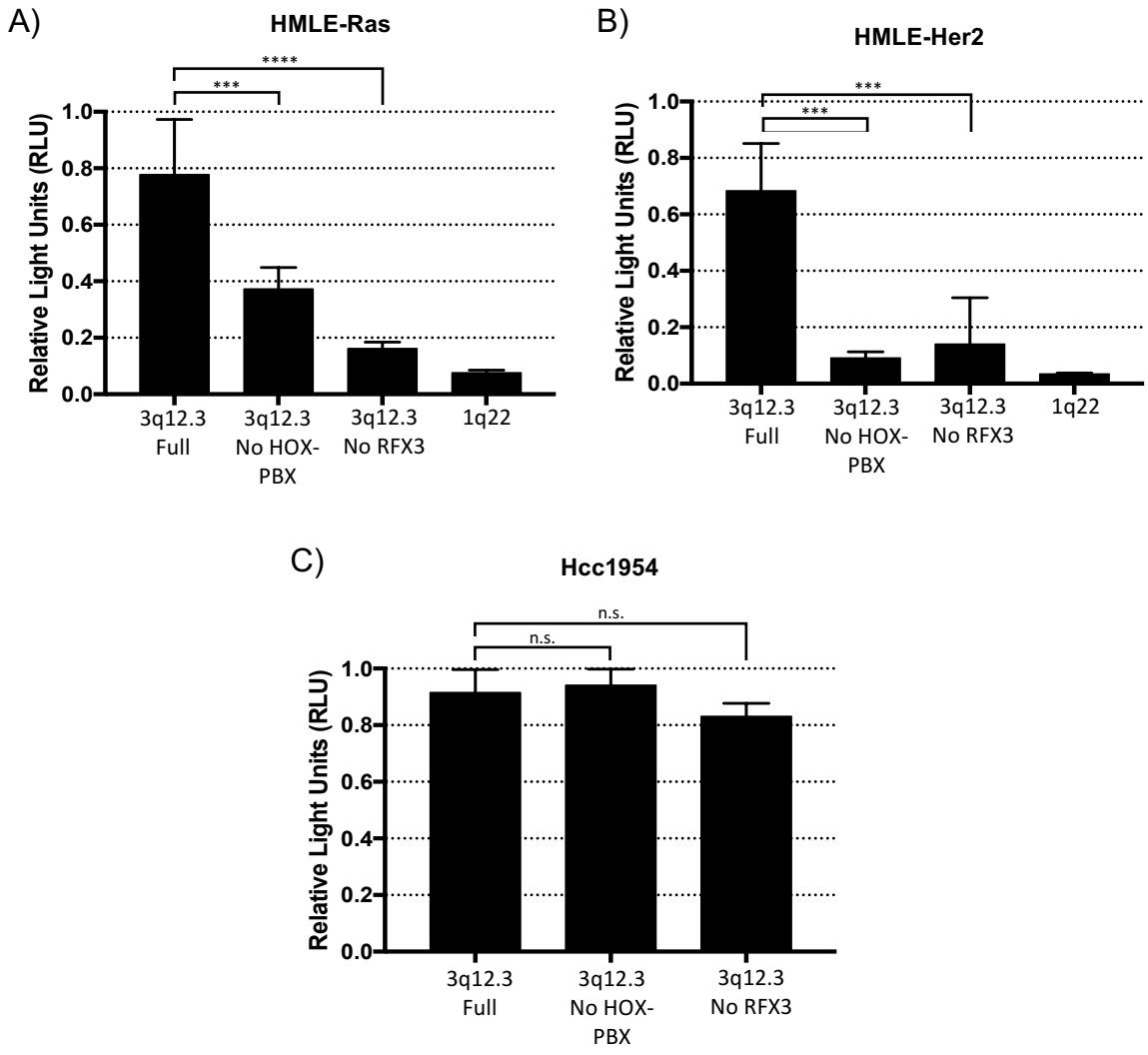


Figure 5-8. Back mutation of critical binding sites to consensus sequences on the 3q12.3 5' LTR results in decreased promoter activity in HMLE cell lines.
Relative 5' LTR promoter activity in (**A**) HMLE-Ras, (**B**) HMLE-Her2, and (**C**) Hcc1954 cell lines. Constructs used either contained full HOX-PBX and RFX3 binding sites or had a binding site removed through back mutation to the consensus sequence. Promoter activity of the 1q22 5' LTR is shown for comparison. Promoter activity is determined as relative light units (RLU) normalized against the internal control *Renilla* expression. Statistical significance was generated by ANOVA with Bonferroni's multiple comparisons test (n.s. = not significant, ***$p<0.0005$, ****$p<0.0001$). All luciferase experiments were conducted in triplicate and data display the mean ± standard deviation.

10A). The RORA binding site differed by eleven nucleotides from the consensus sequence: five transitions, two transversions, one single nucleotide deletion, and one trinucleotide deletion (**Figure 5-9, bottom**). Mutating the RORA site back to the non-active form decreased promoter activity by 5-fold in HMLE-Ras cells. Again, this decreased activity to levels comparable with 1q22 (**Figure 5-10A**). No decrease in promoter activity was seen in the Hcc1954 cell line (**Figure 5-10B**). This cell line had elevated expression of transcription factors known to bind to five unique 11p15.4 sites, including ATF, HIVEP1, IRF5, IKZF5, and RORA (**Figure 5-5, right**).

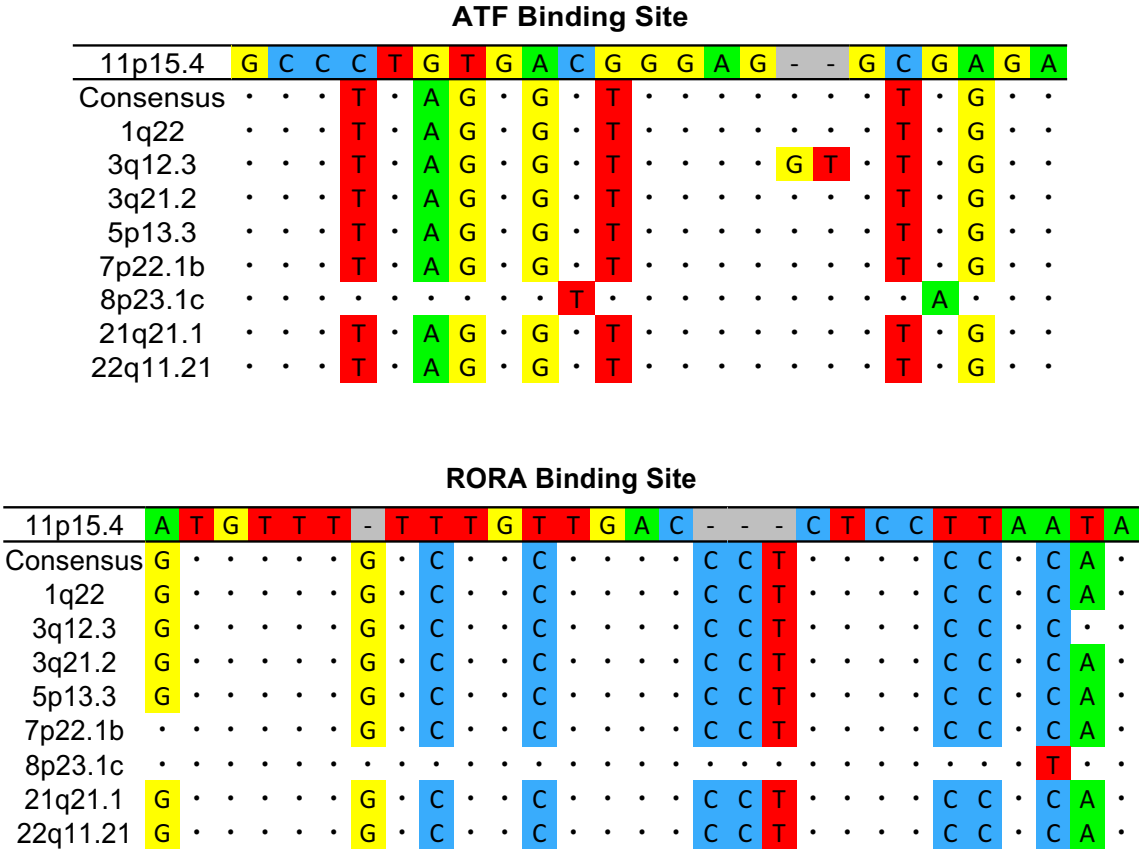


Figure 5-9. Multiple sequence alignments of the ATF and RORA binding sites on the 11p15.4 5' LTR.
Multiple sequence alignment of the ATF (top) and RORA (bottom) binding regions on the nine 5' LTRs of interest in this study as well as a consensus sequence of the site. Sequences are compared against the 11p15.4 5' and 3' LTR sites, with dots used for shared identity and dashes used to indicate indels.

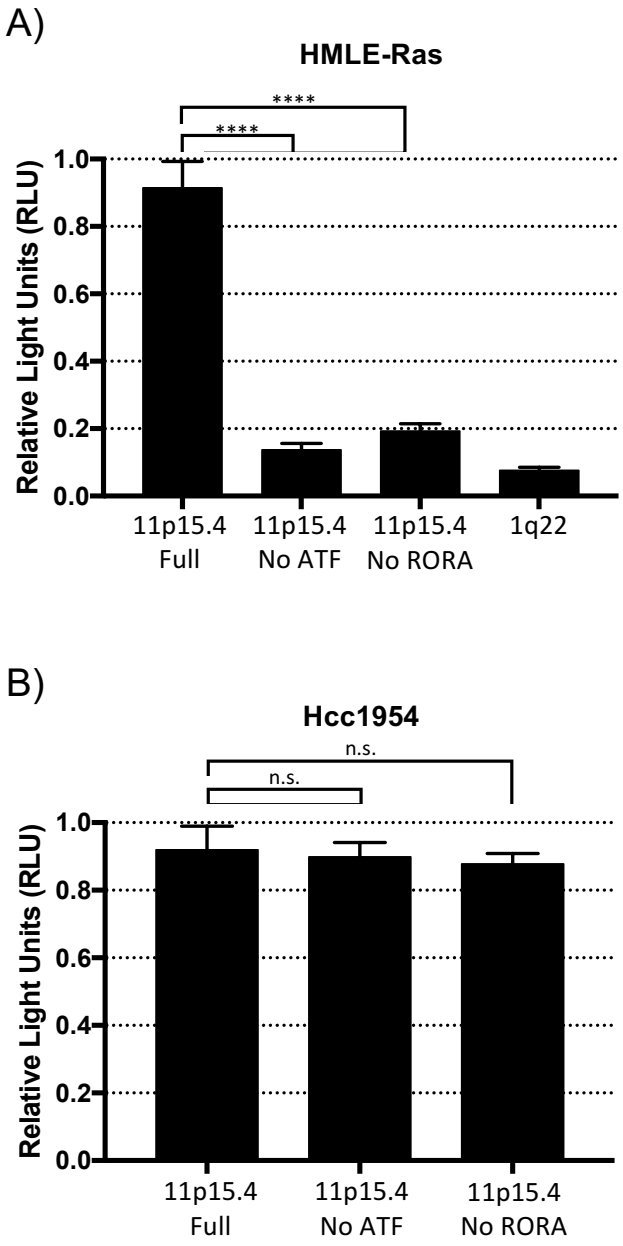


Figure 5-10. Back mutation of critical binding sites to consensus sequences on the 11p15.4 5' LTR results in decreased promoter activity in HMLE cell lines.
Relative 5' LTR promoter activity in (**A**) HMLE-Ras and (**B**) Hcc1954 cell lines. Constructs used either contained full ATF and RORA binding sites or had a binding site removed through back mutation to the consensus sequence. Promoter activity of the 1q22 5' LTR is shown for comparison. Promoter activity is determined as relative light units (RLU) normalized against the internal control *Renilla* expression. Statistical significance was generated by ANOVA with Bonferroni's multiple comparisons test (n.s. = not significant, ****$p<0.0001$). All luciferase experiments were conducted in triplicate and data display the mean ± standard deviation.

5.4 Analysis of unique binding site acquisition and fixation in the human population

Overall, the reporter construct assay results suggested that the HOX-PBX, RFX3, ATF, and RORA binding sites were critical for 3q12.3 and 11p15.4 LTR promoter activity in HMLE cell lines. We decided to analyze these sequences further, by determining if they were acquired over time or if they were present in the viral genome at the time of integration. We additionally analyzed whether they are fixed or polymorphic within the current human population.

Polymorphic proviruses have been investigated previously as possible factors in disease progression. They are believed to be the most likely candidates for contributing to pathogenesis since any provirus with a negative effect on the host would have been selected against over the course of evolution and have never reached fixation (3, 43, 99). Over 10% of all known HML-2 proviruses are either insertionally or allelically polymorphic, however no polymorphic HML-2 element to date has been found to be directly connected with a disease phenotype (41, 99). In spite of these past investigations, information regarding polymorphic transcription factor binding sites within fixed proviruses, and their possible contribution to human disease, is currently lacking.

At the time of integration, the 5' and 3' LTRs of a provirus are generally identical (**Figure 1-3**). Over time, as mutations are accumulated, the sequence variation between the two LTRs increases. The number of nucleotide differences between the two LTRs is often used to estimate the time since integration, but it can also provide information regarding whether a given binding site has been acquired over time. By aligning the 5' and 3' LTRs of 3q12.3 and 11p15.4, we were able to determine if these critical transcription factor binding sites were present at the time of insertion (as evidenced by

their presence in both LTRs) or acquired over time (and found in only one LTR). Using this technique, we determined that one binding site, RFX3, was found in both LTRs of the 3q12.3 provirus and therefore likely to have been present at the time of retroviral insertion. However, three of the binding sites (HOX-PBX, ATF, and RORA), were present in only the 5' LTR and presumed to have been acquired over time (**Table 5-3**). A multiple sequence alignment of all known 5' and 3' HML-2 LTRs was performed in MEGA. Genomic analysis of these three binding sites in the multiple sequence alignment suggested that they were likely acquired by the 5' LTR, rather than lost from the 3' LTR. The three binding sites were created by point mutations or indels specific to the 5' LTR that were not found in most, if any, other known LTRs (**Figures 5-7, 5-9**).

We analyzed the remaining unique binding sites from our original list of 63 sites (**Table 5-2**), with the exception of those found on 7p22.1b and 21q21.1, as these proviruses do not have full 3' LTRs (40). Overall, only 21% (12/56) of the unique binding sites analyzed were found on both LTRs and believed to be present at the time of insertion (**Figure 5-11A**). Most interesting, the majority of these sites (58%, 7/12) were found on the 3q12.3 5' LTR (**Figure 5-11B**).

Table 5-3. Characterization of transcription factor binding sites critical for 3q12.3 and 11p15.4 promoter activity in HMLE cell lines.

Provirus	Binding Site	LTR	Binding Site Allele Frequency	Binding Site Evolution
3q12.3	HOX-PBX	5'	99.68% (fixed)	Acquired
	RFX3	5' and 3'	99.96% (fixed)	Present during insertion
11p15.4	ATF	5'	99.88% (fixed)	Acquired
	RORA	5'	50.76% (polymorphic)	Acquired

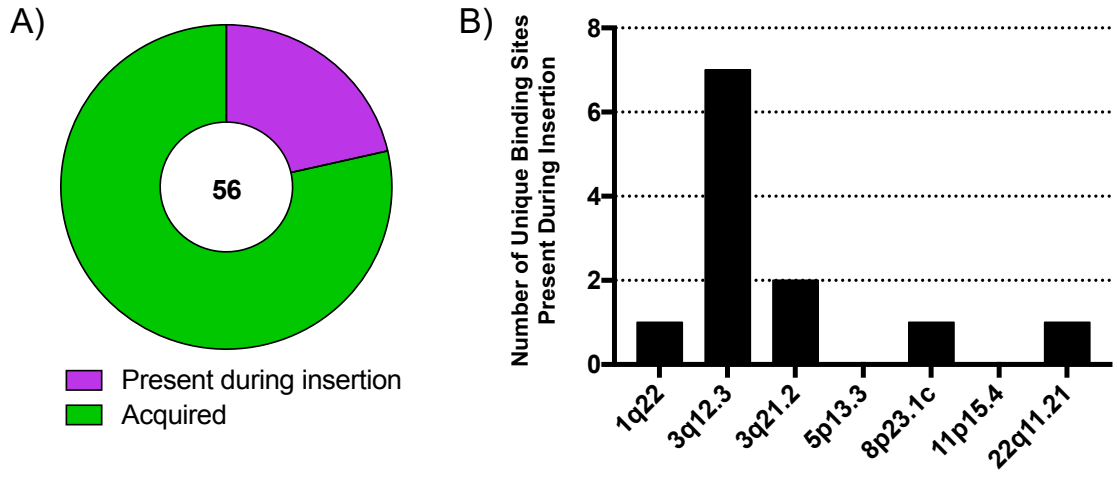


Figure 5-11. Most unique binding sites were not present at the time of retroviral integration.
(**A**) Pie chart showing the number of unique binding sites that were either present at the time of insertion (purple) or acquired over time (green) based on sequence alignments between the 5' and 3' LTRs of the nine proviruses of interest in this study. (**B**) Bar graph showing the number of unique transcription factor binding sites identified as being present at the time of insertion found on each 5' LTR. Binding sites found on the 7p22.1b and 21q21.1 proviruses were excluded from both analyses, as they do not possess full-length 3' LTRs.

In total, 42 binding sites were found in only one LTR and believed to be acquired over time. We again analyzed our multiple sequence alignment containing all known 5' and 3' HML-2 LTRs to determine if these binding sites were acquired by the 5' LTR or lost by the 3' LTR (**Figure 5-12A**). We found that 64% of the binding sites (27/42) were acquired by the 5' LTR, based on the observation of mutations characteristic to the 5' LTR and not found in most, or any, other HML-2 LTR. Only three binding sites appear to be due to a loss of the corresponding 3' LTR sequence. Twelve sites show mutations characteristic of both the 5' and 3' LTR, suggesting that a combination of 5' LTR acquisition and 3' LTR sequence loss resulted in the unique site. Interestingly, 75%

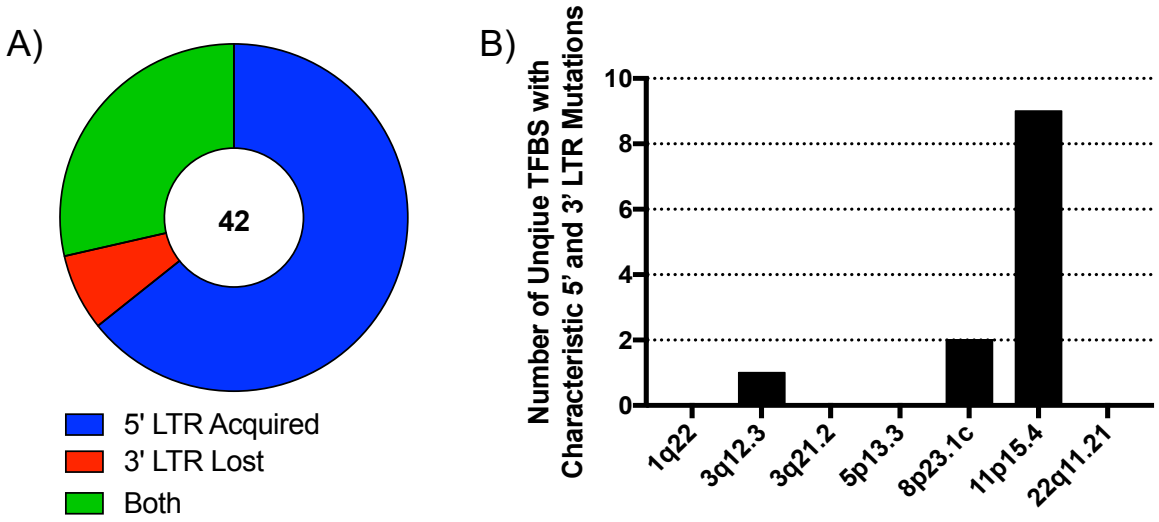


Figure 5-12. Most unique binding sites are acquired by 5' LTR mutations. (**A**) Pie chart showing the number of unique binding sites, not present at the time of integration, that were formed as the result of being acquired by the 5' LTR (blue), lost by the 3' LTR (red), or both (green). (**B**) Bar graph showing the number of unique transcription factor binding sites on each 5' LTR of this study that are identified as being formed due to a combination of 5' LTR acquisition and 3' LTR sequence loss. Binding sites found on the 7p22.1b and 21q21.1 proviruses were excluded from both analyses, as they do not possess 3' LTRs.

(9/12) of the sites with mutations characteristic to both LTRs were found on the 11p15.4 provirus (**Figure 5-12B**).

When comparing the sequences of each 5' and 3' LTR, we were surprised to find that LTR sequence identity did not always correlate with the number of unique binding sites from our identified list (**Table 5-4**). Notably, the 3q21.2, 3q12.3, and 5p13.3 proviruses, which are all of similar age (about 5-10 million years old) had a similar degree of LTR divergence but varied greatly in the number of unique binding sites on their 5' LTRs. All three proviruses had 27-29 nucleotide differences between their LTRs and LTR sequence identities of 97-98%. However, we identified 2 unique binding sites

Table 5-4. LTR sequence identity is not always correlated with number of unique 5' LTR binding sites[a].

Provirus	Estimated Age (my)[b]	LTR Sequence Identity	Number of Nucleotide Differences Between LTRs	Number of Unique Binding Sites[c]
11p15.4	15.44-27.95	93%	71	20
8p23.1c	15-27.17	93%	69	11
3q21.2	4.8-8.69	98%	29	6
3q12.3	5.51-9.98	97%	27	13
5p13.3	6.32-11.4	97%	27	2
22q11.21	1.84-3.34	99%	8	3
1q22	<2	99%	2	1
7p22.1b	<2	N/A	N/A	3
21q21.1	3.46-6.27	N/A	N/A	4

[a]Table is organized in descending order by the number of nucleotide differences between the 5' and 3' LTRs of each provirus.
[b]As determined by Subramanian *et al*. (40)
[c]As determined by **Table 5-2**.

on the 5' LTR of 5p13.3, 6 on 3q21.2, and 13 on 3q12.3. As expected, the largest number of unique binding sites was seen with 11p15.4 and 8p23.1c, two proviruses that were at least 4-27 million years older than any of the other proviruses in our study (40).

To determine whether any of these sites are polymorphic within the human population, we analyzed the VCF (Variant Call Format) files as supplied by phase 3 of the 1000 Genomes Project. The 1000 Genomes Project is an international effort that began in January 2008 with the goal of sequencing the genomes of 1,000 anonymous individuals. Since then, the sampling size has more than doubled to include whole genome sequencing data from over 2,500 individuals. These individuals are from numerous ethnic groups that fall within one of five global super populations: East Asian, Ad Mixed American, African, European, and South Asian (202).

The VCF files supplied by the 1000 Genomes Project provide alternative allele frequencies for both global and super population data. They also provide specific information regarding the type of variance, including whether it is due to a SNP, indel, or large deletion. Using VCFtools, we were able to computationally sub sample VCF files for particular binding sites to analyze their variance within the sequences of all 2,504 individuals currently available by the 1000 Genomes Project.

Of the four binding sites that we found to be critical for HML-2 promoter expression in HMLE cells, three of them had global allele frequencies >99% and were therefore deemed to be fixed in the human population. The RORA binding site, found in the 11p15.4 5' LTR, was found to be polymorphic in the current human population with an allele frequency of 50.76% (**Table 5-3**). In total, only 8% (5/63) of the unique binding sites that we identified in our LTRs of interest were polymorphic (**Figure 5-13, Table 5-5**).

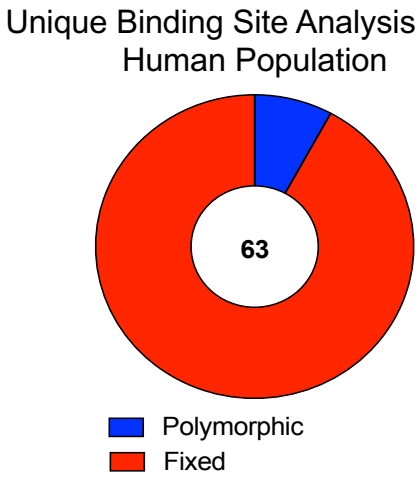


Figure 5-13. Most unique binding sites are fixed in the human population.
Pie chart showing the number of unique binding sites that were either polymorphic (blue) or fixed (red) in the human population based on the genomes of 2,504 individuals from the 1000 Genomes Project. All fixed sites had an allele frequency of ≥89%. All sites that were classified as polymorphic had allele frequencies of ≤52%.

Table 5-5. Characterization of polymorphic HML-2 transcription factor binding sites.

Binding Site	Provirus	LTR	Binding Site Allele Frequency	Binding Site Evolution
HHEX	11p15.4	5’	50.76%	Acquired
IK3	3q21.2	5’	42.41%	Acquired
NRL	11p15.4	5’	52.23%	Acquired
RORA	11p15.4	5’	50.76%	Acquired
TEF	11p15.4	5’	50.16%	Acquired

All five polymorphic binding sites were present only in the 5’ LTR. Interestingly, all but one binding site (IK3), was from the 11p15.4 provirus (**Table 5-5**). Genomic analyses of these binding sites using a multiple sequence alignment of all known 5’ and 3’ HML-2 LTRs indicated that four of them (HHEX, IK3, RORA, and NRL) were acquired by the 5’ LTR. These four sites contain either deletions or SNPs not present in most, or any, other known HML-2 LTR. In contrast, the TEF binding site found on the 11p15.4 5’ LTR appeared to be acquired in part from sequences gained by the 5’ LTR (as evidenced by a G to A transition) as well as sequences lost from the 3’ LTR (including a 1 bp deletion as well as a C to T transition). These mutations were characteristic of each LTR and not found in most other known HML-2 LTRs.

We further analyzed the allele frequencies of the five polymorphic binding sites by super population to see if any were more abundant in a particular ethnic group (**Figure 5-14**). Overall, each binding site had slightly higher allele frequencies in Ad Mixed American and African populations, where 60% of the population or greater possessed the binding site. One exception was the IK3 binding site, which had an allele frequency of

40-50% in all super populations with the exception of the African group. Surprisingly, the allele frequency of IK3 in the African population was only ~20% (**Figure 5-14B**).

We conducted a pairwise linkage analysis of the SNPs responsible for the polymorphic quality of these sites using the Ensembl database and its "linkage disequilibrium" function (203). Our analysis indicated no evidence of linkage disequilibrium between any of the SNPs on the 11p15.4 5' LTR. The polymorphic IK3 site on the 3q21.2 5' LTR comprised two SNPs: rs12631168 and rs2134522. The former (rs12631168) is mostly fixed, with an estimated 89% of the population containing an A at this site and 11% possessing a T. The latter SNP (rs2134522) is highly polymorphic; 46% of the population has an A at this site while the remaining 54% has a C. Ensembl analysis indicated that these sites exhibit low levels of linkage disequilibrium, with an average r^2 value of about 0.1 (0 signifying complete linkage equilibrium and 1 signifying complete coinheritance).

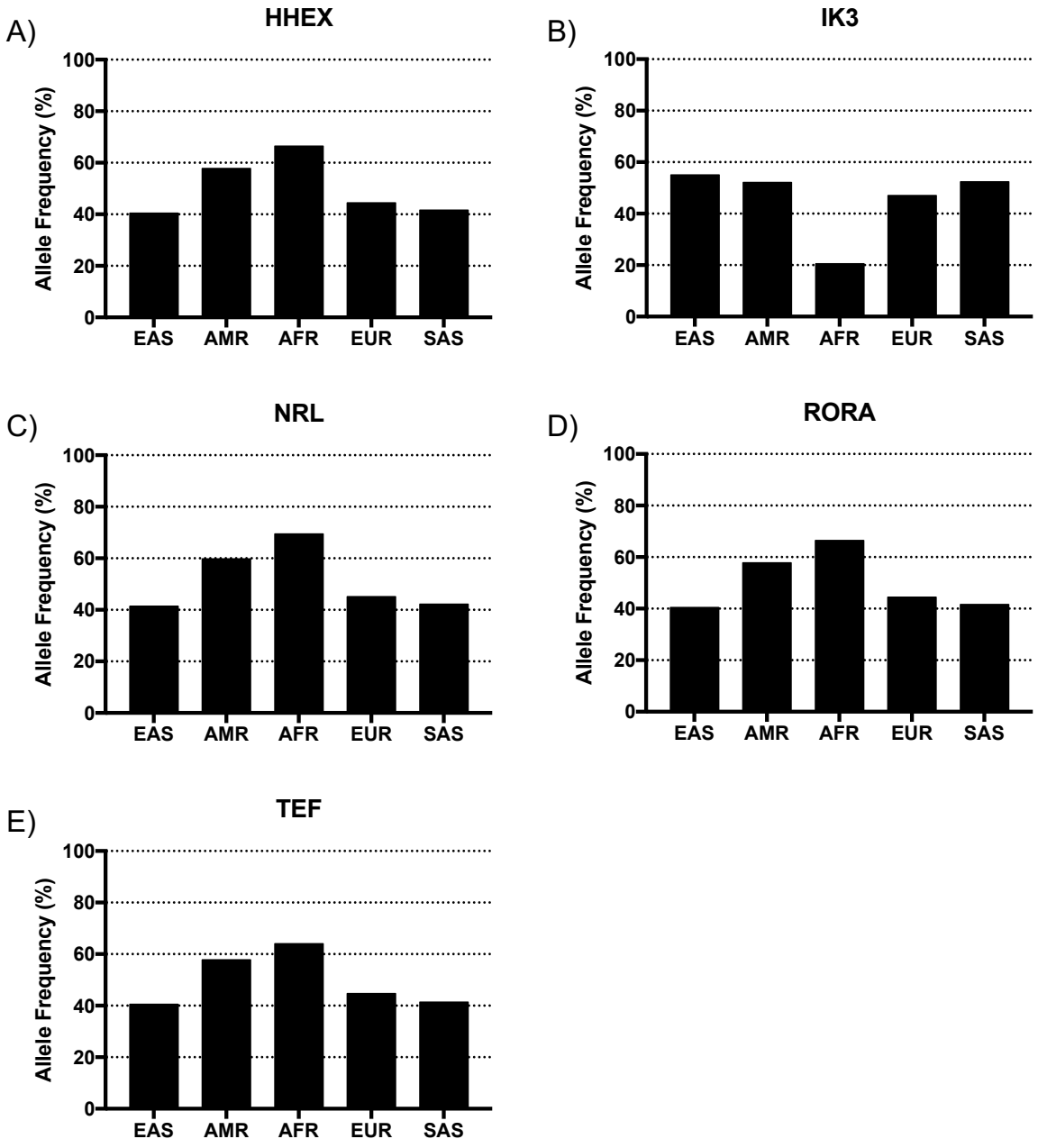


Figure 5-14. Allele frequencies of polymorphic HML-2 5' LTR transcription factor binding sites within each super population.
Allele frequencies were determined from 2,504 individuals from the 1000 Genomes Project and broken down by super population for each of the five polymorphic binding sites identified: (**A**) HHEX, (**B**) IK3, (**C**) NRL, (**D**) RORA, and (**E**) TEF. EAS, East Asian; AMR, Ad Mixed American; AFR, African; EUR, European; SAS, South Asian.

5.5 Evolution of the HOX-PBX and RORA binding sites in HML-2 LTRs

Alignment of the 3q12.3 5' and 3' LTRs showed a 4 bp insertion, found in the middle of the HOX-PBX site, that appeared to be a GATT sequence duplication (**Figure 5-7, top**). This provirus is estimated to have integrated ~5-10 million years ago and can be traced back through the primate lineage to gorillas (40). Using the UCSC Genome Browser, we examined this LTR in several non-human primate reference genomes. We found that although the 3q12.3 provirus is conserved across multiple species (including humans, chimpanzees, and gorillas), the 4 bp insertion, and consequently the HOX-PBX binding site, is only present in the sequences of Denisovans and modern humans (**Figure 5-15**). These results suggest that this binding site was acquired very shortly after the human-chimpanzee evolutionary split and that it has been stably integrated in the human genome ever since (**Table 5-3**).

Denisovans, as well as Neanderthals, are extinct hominin species, evolutionary distinct from one another. Fossil records indicate that these two species pre-date the emergence of anatomically modern humans and existed unitl at least 30,000 years ago.

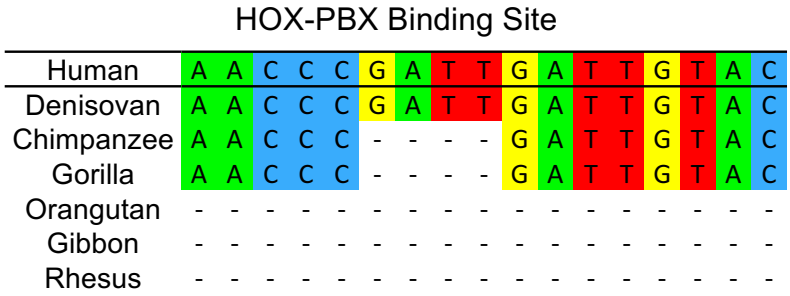


Figure 5-15. Evolution of the HOX-PBX binding site.
Multiple sequence alignment of the human HOX-PBX binding site as compared to the homologous sequences in different non-human primate reference genomes. Dashes indicate indels.

Traces of Neanderthal DNA can be found in all modern non-African human populations while Denisovan DNA has only been detected in Oceanic and Southeast Asian populations (235, 236). Although we searched for the presence of the HOX-PBX binding site in the Neanderthal reference genome as supplied by the UCSC Genome Browser (112), read coverage in this area was not sufficient to confirm its existence.

The RORA binding site on 11p15.4 was one of only five polymorphic binding sites that we identified. This polymorphism is due to a single nucleotide change, where 51% of the human population contains an A at the 23rd base pair in the site (and therefore an intact RORA site) and 49% of the population contains a T (and an inactive binding site). This provirus is of particular interest because 11p15.4 is a segmental duplication of 8p23.1c, the latter of which is estimated to have integrated ~15-27 million years ago in the ancestral primate genome. Although the proviral sequence is quite old, the duplication onto human chromosome 11 occurred after the human-chimpanzee split, and the 11p15.4 sequence is human-specific (40). We aligned the 5' and 3' LTRs of these two proviruses and compared the sequences at the sites homologous to RORA. We found that although the 3' LTRs at this site were identical, the 5' LTRs differed by only one nucleotide, at the same 23rd base pair position that is responsible for the RORA polymorphism (**Figure 5-16**). Based on these observations, it appears as though half of the human population has evolved away from the ancestral 8p23.1c sequence and acquired a SNP that results in an active RORA binding site.

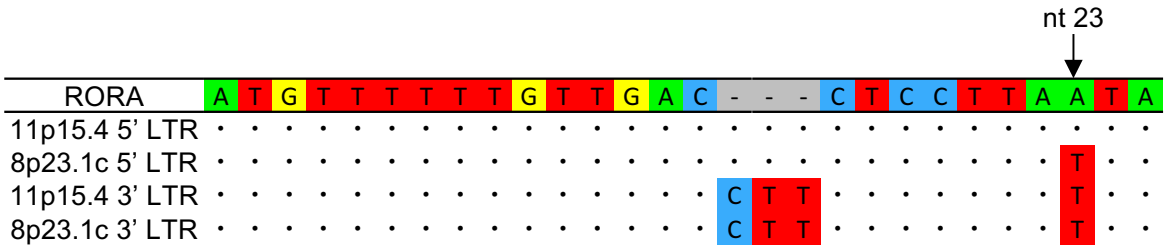


Figure 5-16. Evolution of the RORA binding site.
Multiple sequence alignment of the RORA binding site as compared to the 5' and 3' LTRs of the 11p15.4 and 8p23.1c proviruses. Dots are used for shared identity and dashes indicate indels. The 23rd nucleotide in the sequence, which appears to be responsible for the RORA polymorphism, is shown with an arrow.

Chapter 6: Discussion and Future Directions

6.1 Discussion

HERV-K (HML-2) proviruses are expressed at low levels in many non-diseased cells but upregulated in certain human malignancies (3, 92, 98). However, the regulation, mechanism, and consequences of ERV activity are poorly understood. HML-2 expression appears to be dependent upon an assortment of factors, including epigenetics, proximity to transcribed host genes, and transcription factor binding (91, 110, 162, 237). We chose to first explore the latter component by investigating the interplay between *trans*-activating factors and *cis*-regulatory sequences located on the 5' LTR.

We conducted a series of dual-luciferase assays across a panel of eighteen human cell lines, testing the retention of 5' LTR promoter activity of nine HML-2 proviruses. Our goal was to elucidate whether LTRs of similar sequence also exhibit similar patterns of promoter expression. We detected negligible LTR activity in our two immortalized HME control cell lines and a significant upregulation of promoter expression in 73% (11/15) of the tumorigenic breast cancer cell lines (**Figure 3-3**). These results are in accordance with previously published literature that suggests around 75-85% of breast cancer samples exhibit increased HML-2 expression (50, 120, 126).

The molecular subtype of each breast cancer cell line was noted (**Table 3-3**) but, consistent with the preliminary single-genome sequencing results conducted by Ravi Subramanian (**Table 3-1**), no correlation between hormone receptor status and level of HML-2 expression was seen (**Figure 3-5**). These results are in contrast with a recent report suggesting that HML-2 expression is strongly correlated with the basal-like molecular subtype (238). That study performed a meta-analysis of RNA-Seq files from

The Cancer Genome Atlas (TCGA) database. Although the sample size from TCGA was substantial (512 breast cancer patients), the study only examined the transcript level of four HML-2 proviruses: 7p22.1, 6q14.1, 19p12b, and 8p23.1. As there are over 90 HML-2 proviruses in the human genome, more elements should be taken into consideration before definitively concluding that HML-2 expression is correlated with a particular molecular subtype.

We performed pairwise comparisons between LTR sequence identity and patterns of significant expression within our cell line panel and found the two conditions to be positively correlated in nearly all instances (**Figure 3-6**). Our results indicated that LTRs with ~70% identity shared promoter expression patterns ~60% of the time, whereas LTRs with ~95% sequence identity shared promoter expression patterns ~90% of the time. Overall, these results suggest that HML-2 LTRs with similar sequences share comparable patterns of promoter activity, likely due to the conservation of similar *cis*-regulatory sequences.

Dual-luciferase assays are excellent molecular tools for analyzing the retention of promoter activity in specific cell lines under optimal epigenetic conditions. However, because they involve transient transfections, they do not take all endogenous cellular activities into consideration such as epigenetic or chromatin modifications, enhancer-mediated effects from nearby elements, or RNA interference (206, 207). These factors are important to consider since previous studies demonstrate increased HML-2 transcription from proviruses that no longer possess a 5' LTR, suggesting that at least some proviruses are transcribed in an LTR-independent manner (87, 110, 239). These results were also seen in the preliminary single-genome sequencing study, where the

16p11.2 provirus exhibited one of the highest levels of transcription production, despite not having a 5' LTR (**Table 3-1**) (40).

Previous investigations into the study of HML-2 transcription relied upon techniques such as DNA hybridization, RT-PCR, amplicon sequencing and single-genome sequencing. However, these methods are not stringent enough to accurately catalog provirus transcription. They rely upon consensus sequences, often utilizing LTR- or gene-specific primers (38, 50, 105). Proviruses that contain indels or a large number of point mutations, such as those inflicted by APOBEC3G proteins, may not possess the consensus sequences used to create the hybridization probes or primers, and would likely go undetected. Because sequence similarity of HML-2 proviruses can be over 99%, these techniques cannot distinguish between individual elements and primer bias in PCR amplification can encumber results (40, 156). Furthermore, orientation of transcription, an important factor to consider when investigating potentially functional ORFs and HML-2 protein production, would be overlooked unless strand-specific primers were used.

We opted to bypass the abovementioned limitations by characterizing HML-2 transcription via RNA-Seq, an NGS technique that reliably detects a wide variety of transcripts (including mRNAs, lncRNAs, and miRNAs), allowing one to determine the abundance level of a transcript of interest as well as its orientation of transcription, splice patterns, transcription start sites, and any read-through transcription into downstream genes. Most current RNA-Seq platforms and methods of analysis rely upon the alignment of very short reads, often 50-100 bp in length, to the human reference genome (182). This methodology creates an issue when investigating the transcription of repetitive elements,

which are too similar to be differentiated by stretches of 100 bp or less and are often annotated incorrectly within the reference. At least 36 HML-2 elements are known to be either absent or incorrectly annotated in the hg19 build of the human reference genome. Repetitive elements are often incorrectly discarded during genome assembly and non-reference HML-2 proviruses may also be due, in part, to polymorphic integrations not present in the individual(s) used to create hg19 (41, 183).

Alongside a previous graduate student in our lab, Neeru Bhardwaj, I helped produce and validate an RNA-Seq analysis pipeline that is optimized for the capture of HML-2 expression. This protocol utilized Illumina MiSeq sequencing and relied upon long (301 bp), paired-end reads, produced from stranded libraries, that were aligned to both hg19 and a custom HML-2 reference genome containing all known HML-2 elements. These reads were filtered for unique alignments only, ensuring accurate assessment of proviral expression. This protocol was validated in Tera-1 cells, a human teratocarcinoma cell line known to produce high levels of HERV-K (HML-2). Importantly, this pipeline analysis proved successful at identifying transcript abundance levels from polymorphic proviruses not annotated in hg19 as well as all young proviruses with high sequence similarity to one another (110).

We adopted the same sequencing protocol that was used to analyze HML-2 transcription in Tera-1 cells to the current study, with intent to characterize the HML-2 transcriptome during human mammary epithelial cell transformation. We sequenced one immortalized, non-diseased cell line (HME) as well as three tumorigenic cell lines (HMLE-Ras, HMLE-Her2, and Hcc1954). For our analysis, we chose to focus on full-length proviruses only, since these elements have the potential to encode protein that may

have functional consequences for the tumor cell, be useful in immunotherapy, or be utilized as a molecular biomarker. Solo LTRs, which are present in the human genome at a rate 10-fold higher than full-length proviruses, may also play a role in altering host gene expression through the donation of alternative promoters, enhancers, splice sites, and termination signals, and warrant a separate, future investigation.

Filtering for uniquely mapped reads only is critical for maintaining alignment specificity and being able to accurately determine the identity of provirus expression. Our alignment parameters allowed for up to two mismatches per read, in the event that there was a low level of SNPs in the donor sequence that were not present in the reference. Overall, our data had a 96-99% retention rate after filtering for unique alignments (**Figure 4-4**). This high retention rate was partly due to the stringency of our protocol, which was optimized for read length and alignment specificity, as well as the individual proviruses that were expressed. Although the biological activity of HERV-K (HML-2) is often credited to its young age relative to other HERVs, we found that all but one expressed provirus (1q22) had an estimated age of at least 5 million years old (**Figure 4-9**). Since older proviruses have accumulated, on average, more mutations and sequence diversity than younger ones, it was not surprising that we saw low rates of multi-mapping. The reads that did multi-align were short, averaging 112 bp in length, and were mostly located to the LTR regions of proviruses with solo LTRs of highly similar sequence. These observations support the utilization of longer read lengths when analyzing HERV transcription.

HML-2 expression was detected in every cell line sequenced, with highest levels seen in the tumorigenic Hcc1954 and HMLE-Ras cell lines. Total expression levels in

these two cell lines was about half that exhibited by Tera-1 cells (**Figure 4-5**). These results are reflected in our original luciferase panel, which showed total promoter activity in the Tera-1 cell line being 2-fold higher than T47D, the breast cancer cell line with the highest combined HML-2 LTR activity (**Figure 3-3**). Surprisingly, we did not see increased HML-2 expression in the HMLE-Her2 cell line and this was the only tumorigenic cell line that did not exhibit LTR-driven expression (**Table 4-3**). It is possible that the overexpression of *ERBB2* resulted in a transcriptional environment, different from the other cell lines, that does not support increased HML-2 activity. Although previous work suggests up to 85% of breast cancer samples have upregulated HML-2 expression, it is not found in all cases and these cells will be subject to future investigations (50, 120, 126).

HML-2 expression in breast cancer cell lines has been shown to drastically increase in response to treatment with female steroid hormones, specifically estradiol and progesterone (105, 130, 240). However, we did not test for hormone responsiveness in our study. Due to the absence of expression of the three major hormone receptors (estrogen, progesterone, and HER2) in most of the cell lines that we analyzed (**Figure 4-2**), it is unlikely that induced hormone expression would have had any effect on HML-2 transcription in this study.

Our HML-2 reference genome contained all 36 currently known non-reference HML-2 elements. It was important to include polymorphic proviruses in our investigation since they are the main candidates for contributing to disease, as it is believed that any provirus with deleterious effects on the host would have been selected against and removed from the population over the course of evolution, preventing fixation (3, 43, 99).

This list includes the recently isolated polymorphic provirus, Xq21.33. This integration is rare, with an estimated allele frequency of <2%, but appears to possess all intact ORFs and no obviously defective mutations (41). However, the replication competency of this provirus is still under investigation.

Our analysis found significant expression from 15 HML-2 proviruses in the four cell lines sequenced (**Figure 4-6**). Parallel alignments to both hg19 and HML-2 reference genomes indicated no significant discrepancies (**Figure 4-5**), suggesting no expression from non-reference HML-2 proviruses in this study. Results were separately analyzed for transcription occurring in sense vs. antisense orientation. Between 49-82% of total HML-2 expression was due to antisense transcription from an assortment of proviruses (**Figure 4-8**). In contrast, the majority of sense-transcribed HML-2 transcription stemmed from only three loci: 1q21.3, 1q22, and 3q12.3 (**Figure 4-7**). Two of these proviruses (1q22 and 3q12.3) were originally identified as the highest expressing proviruses in our preliminary single-genome sequencing analysis (**Table 3-1**) and 3q12.3 exhibited high promoter activity in nearly every neoplastic cell line during our luciferase experiments (**Figure 3-3**). 1q22 expression was detected from all three HME and HMLE cell lines, but not from the Hcc1954 cell line. IGV analysis of reads aligning to this region indicate that transcription is not driven by the native 5' LTR, but instead from an upstream simple repeat element as part of an annotated lncRNA (112, 184, 231). It is possible that this lncRNA, and therefore 1q22 expression, is specific to the individual from which the HMECs were originally isolated.

ORF analysis of the sense-transcribed proviruses in this study did not show coding capacity for any of the main proviral genes (*gag, pro, pol,* and *env*) (40).

Although the accessory proteins Rec and Np9 have been documented in both tumorigenic and healthy cells (241), no subgenomic transcripts necessary for their production were detected in our data set. Rec and Np9 activity have also been shown to be associated with increased expression of the c-MYC proto-oncogene through inhibition of the PLZF protein (148-150). However, no correlation between HML-2 expression and c-*myc* transcription was observed in any cell line examined (data not shown). These results suggest that HML-2 accessory genes did not play a role in our transformation model.

Due to the age of the expressed proviruses in this study, we did not expect may to still have functional LTRs. Furthermore, analysis of the genomic regions surrounding each expressed provirus showed that most of them were integrated near or within genic regions, suggesting that nearby host gene transcription may be influencing proviral expression. Indeed, we found that the majority of expression, including all antisense transcripts, was due to either intronic or read-through transcription (**Figure 4-10**). IGV analysis of these two modes of transcription suggest that these proviruses are not self-transcribed from their native LTRs. Instead, their expression is either seen as a consequence of being preserved within an incompletely removed intron or from being situated downstream of a highly transcribed host gene or repetitive element.

The frequent use of intronic or read-through transcription may explain previous studies that demonstrated upregulation of transcripts from proviruses that no longer have 5' LTRs (87, 110, 239). Interestingly, almost all proviruses exhibiting antisense transcription in this study had intact 5' LTRs. Exceptions were seen with 7q34, 14q11.2, and 1q21.3, all of which possess substantial deletions and nearly complete loss of the 5' LTR sequence (**Table 4-3**) (40). Although evidence suggests that LTRs can be

bidirectional and that 3' LTRs are capable of producing antisense transcripts (87, 242, 243), we did not see any transcripts originating from the 3' LTR that were consistent with such activity.

In addition to intronic and read-through transcription, HML-2 sense transcripts also appeared to originate from lncRNA-associated and LTR-driven modes of transcription (**Figure 4-10**). In total, we detected three instances of active HML-2 LTR use. The 1q22 3' LTR was responsible for the donation its polyadenylation site found at its R-U5 border for termination of an annotated lncRNA (BC041646) of unknown function. This lncRNA is multi-spliced and the 1q22 5' LTR, being located in the second intron, is ultimately excised from the transcript and not used for promotion. Instead, transcription is initiated from a (TCC)n simple repeat element located about 12.7 kb upstream from the 1q22 provirus (112, 184, 231).

Although there is increasing belief that antisense and/or lncRNA-associated ERV transcripts may play key roles in gene expression, their impact on host gene regulation is currently unclear and only a few instances have been documented. Recently, a HERV-H LTR was found to function as an enhancer for lncRNA production in human embryonic stem cells (hESC), and play a key role in maintaining hESC pluripotency (56, 181). Additionally, a novel lncRNA, produced by a MER38 LTR element, was recently suggested to exhibit tumor-specific activation in a number of cancer types including pancreatic, lung, rectal, colon, and stomach (244). BC041646 transcription did not exhibit disease-specific expression patterns in our study since it was detected in both non-diseased (HME) and tumorigenic (HMLE) cell lines. As mentioned previously, the lack of BC041646 expression in the highly active Tera-1 and Hcc1954 cell lines suggest that

the lncRNA expression may be specific to the individual from which the HME and HMLE cell lines were derived, as opposed to being disease-specific. However, the observation that ERVs can donate regulatory elements, such as polyadenylation sites, for lncRNA production gives further credence towards the importance of studying their patterns of expression and regulation in both non-diseased and malignant cells.

There were two instances of active 5' LTR promoter use in the cell lines sequenced. The 3q12.3 provirus appeared to have LTR-driven transcription in the Hcc1954 and HMLE-Ras cell lines while the 4p16.1b provirus appeared to have LTR-driven transcription in only Hcc1954. 5' LTR promoter activity for these two proviruses was confirmed through a dual-luciferase assay (**Figure 4-13**). The 4p16.1b 5' LTR exhibited low levels of promoter activity in only the Hcc1954 cell line, consistent with the low level of transcription found in only these cells. The 3q12.3 5' LTR exhibited high levels of promoter expression in all three tumorigenic cell lines. Interestingly, transcripts originating from this provirus were only seen in the Hcc1954 and HMLE-Ras cell lines. Since luciferase assays are based on transient transfections and are insensitive to epigenetic effects, it is possible that 3q12.3 was silenced endogenously in the HMLE-Her2 cell line, possibly by CpG methylation, but capable of promoter activity under optimal epigenetic conditions.

Genomic analysis of the 4p16.1b locus showed no known host genes within 100 kb of the provirus. In contrast, seven host genes were found in close proximity to the 3q12.3 provirus. However, despite evidence suggesting that retroviral LTRs can influence host gene transcription up to 100 kb away from their site of integration (3, 4),

we found no correlation between provirus expression and host gene transcription in any cell line sequenced (**Figure 4-14**).

LTR-driven transcription was the only mechanism found to be dependent upon the tumorigenicity of the cell, as no evidence of 5' LTR activity was detected in the non-diseased cells. This observation in combination with the previous luciferase panel results suggest a similar proposition: LTR promoter activity, and therefore LTR-driven transcription, is reliant upon tumorigenic transcriptional environments. Additionally, LTR sequences, likely *cis*-regulatory sequences found in the U3 region, appeared to play a large role in initiating proviral transcription. Exactly what cellular factors are required for LTR promoter activity became the next phase in our investigation.

Research suggests that HML-2 provirus transcription is TATA- and Inr-independent and that activity is regulated by transcription factor binding (162, 204). Luciferase experiments using truncated versions of 22q11.23 LTR-Hs, an HML-2 element with a very high transcript abundance level in Tera-1 cells, showed that promoter activity is not significantly altered after removal of both the R and U5 regions. Instead, our results suggested that LTR activity is dependent upon *cis*-regulatory sequences located in the 3' half of the U3 region (**Figure 5-3**). We used these results to identify which U3 binding sites are critical for driving promoter expression in our HMEC transformation model.

To identify critical binding sites, we returned back to the original nine 5' LTRs used in our luciferase experiments. Of those nine LTRs, only two of them were significantly upregulated in HMLE cells as compared to HME cells; 3q.12.3 had increased activity in both HMLE-Ras and HMLE-Her2 whereas 11p15.4 had increased

activity in only the HMLE-Ras cell line (**Figure 5-4**). We sought to investigate which binding sites on these 5' LTRs were necessary for promoter activation in HMLE cells.

Using a transcription factor binding site prediction algorithm (201), we identified 63 unique binding sites across these nine LTRs. Each binding site was considered "unique" as compared to the other eight 5' LTRs. In total, 22 of these binding sites were unique amongst all known HML-2 elements, including all other proviruses and solo LTRs, 26 were not found in any other proviral LTR but were found in at least one solo LTR, and 15 were found in at least one other provirus (**Table 5-2**). 3q12.3 and 11p15.4, the two proviruses with the highest levels of promoter activity (**Figure 3-3**), had the most unique sites. 3q12.3 had 13 unique sites, 7 of which were not found in any other HML-2 element. 11p15.4, despite being a segmental duplication of 8p23.1c, had 20 unique sites, 13 of which were not present in any other HML-2 element.

Comparisons between LTR sequence divergence and number of unique binding sites identified through our analyses showed that these two conditions are not always correlated (**Table 5-4**). However, our results suggest that older proviruses have accumulated, on average, more mutations and therefore more unique binding sites. Despite previous notions that the younger, more intact proviruses have retained the most biological activity, we postulate that older proviruses should not be discredited as noteworthy contributors to HML-2 activity. Older proviruses, with more unique transcription factor binding sites, may contribute to the variant expression often seen during human disease.

We used our RNA-Seq results to investigate whether any transcription factors, known to bind to the unique sites on the 3q12.3 and 11p15.4 5' LTRs, were upregulated

in the HMLE cell lines as compared to the HME cell line. In particular, we were looking for candidate binding sites that may explain the upregulation of promoter activity seen in HMLE-Ras and HMLE-Her2 cells as compared to the non-diseased cells. These results provided us with two candidate binding sites per 5' LTR: the HOX-PBX and RFX3 sites on 3q12.3 and the ATF and RORA sites on 11p15.4 (**Figure 5-5**). All four of these binding sites are unique to each provirus and are not found in any other known HML-2 element. Additionally, they are all located in the 3' half of the U3 region (**Figure 5-6**), the same area that was proposed to be critical for promoter activity in our previous Tera-1 experiments (**Figure 5-3**).

Individual removal of these sites decreased LTR promoter activity in HMLE-Ras and HMLE-Her2 cells by 2- to 7-fold, suggesting that they are individually critical for LTR promoter activity in HMLE cells. By comparison, removal of these binding sites did not significantly decrease promoter activity in the Hcc1954 cell line, which showed elevated expression of transcription factors known to bind to five unique sites on both the 3q12.3 and 11p15.4 5' LTRs (**Figure 5-8, 5-10**). These results suggest that, if more binding sites are available, removal of one site does not significantly impact LTR promoter activity.

All four of these sites are known to be involved with transcriptional activation, particularly during the regulation of human embryogenesis (245-248). Interestingly, this observation corroborates with previous literature that suggests HERVs are regulated in manners similar to stem cell genes, by relying on cell-specific transcription factors and epigenetic modifications rather than TATA boxes or Inr elements (162, 204). Recent studies have investigated the role of pioneer transcription factors in the regulation of stem

cell genes. Unlike conventional transcription factors, pioneer factors are able to directly bind heterochromatin, open the region to allow for the binding of additional host factors, and passively enhance transcription. Pioneer factors are known to predominantly regulate genes during embryonic development and human cancer (249, 250).

PBX1, a known pioneer factor, has been shown previously to bind to HOX-PBX motifs, the same binding site we found to be critical for 3q12.3 5' LTR promoter activity in HMLE cell lines. PBX1 is a member of the homeodomain family and is traditionally involved with the regulation of organogenesis. PBX proteins are cofactors for HOX transcription factors and their interaction results in increased affinity and specificity for binding to HOX-PBX sites (250, 251). PBX1 is under investigation as a therapeutic target for cancer treatment, as it is known to be overexpressed in a number of different cancers including breast, lung, ovarian, melanoma, and prostate tumors (252-256). Small molecules designed to target the interaction between HOX and PBX proteins are able to decrease the proliferation of tumorigenic cells (254, 257). Additionally, PBX1 is being considered as a therapeutic target for ER+ luminal breast cancers, as studies suggest that PBX1 is a critical factor for ERα-mediated transcription (251). The ability of pioneer transcription factors, including PBX1, to bind hypermethylated LTR sequences and initiate proviral transcription, will be the subject of future investigations.

All four critical binding sites (HOX-PBX, RFX3, RORA, and ATF) were unique to all other known HML-2 elements and we were interested in examining how each site was acquired. By aligning the 5' and 3' LTRs of a provirus, we were able to determine if a given site was present at the time of retroviral integration (and therefore present in both LTRs) or acquired over time (and present in only one LTR). This analysis was based on

the observation that, during the integration step of the retroviral life cycle, both LTRs are identical (8). Therefore, it is presumed that if a binding site is only found in one LTR, it was acquired over time through various mutations.

Only 21% of the unique binding sites that we identified were characterized as being present at the time of insertion. Furthermore, the majority of them (7/12), were located on the 3q12.3 5' LTR (**Figure 5-11**). Included in this list is the RFX3 site, shown to be critical for 3q12.3 promoter activity in HMLE cells. The large number of unique binding sites on the 3q12.3 5' LTR suggests that this particular retrovirus possibly accumulated a substantial amount of genetic drift by the time of its last replication cycle. The high degree of unique sites present at the time of insertion may also explain why the 3q12.3 5' LTR had a widely different expression pattern than the other LTRs in this study (**Figure 3-3**).

The second unique binding site found to be critical for 3q12.3 activity in HMLE cells was HOX-PBX. This binding site is only present in the 5' LTR and deemed to be acquired over time. Although the 3q12.3 provirus can be traced back through the primate lineage to gorillas, the HOX-PBX binding site is only found in the 5' LTR of Denisovans and modern human genomes (**Figure 5-15**). This site, created by a duplication of GATT sequence, appears to have been acquired shortly after the evolutionary split between humans and chimpanzees. Importantly, genomic analysis into the sequences of 2,504 individuals, as supplied by the 1000 Genomes Project, suggests that this binding site has been fixed in the human population ever since, as it has an allele frequency of 99.68% (**Table 5-3**). This observation suggests that although the 3q12.3 provirus is evolutionarily conserved amongst several non-human primate species, the HOX-PBX binding site is

human-specific. HOX proteins are widely expressed during development, but aberrant expression has been documented during malignancy and increased HOX gene expression is being investigated as a potential breast cancer biomarker (258). Additionally, as discussed previously, PBX proteins known to bind to this site are aberrantly expressed in tumor cells and could function as pioneer factors to drive transcription of hypermethylated LTRs (250, 251).

Whether any 5' LTRs have acquired unique binding sites through positive selection for a biological purpose is still unclear, although enticing to consider. Of the 42 unique binding sites in our data that appeared to be acquired over time, the majority of them seem to be acquired by the 5' LTR rather than lost by the 3' LTR (**Figure 5-12A**). Differentiation between 5' LTR acquisition or 3' LTR loss was based on a multiple sequence alignment of all known 5' and 3' HML-2 LTRs, using sequences obtained from the hg19 build of the human reference genome (112, 184). Mutations in the binding site of interest that were characteristic to the 5' LTR, and not found in most or any other HML-2 LTRs, were presumed to be acquired. Only three unique binding sites appeared to be associated with mutations that disrupt the binding site sequence in the 3' LTR. These instances are most likely due to unique binding sites that were present in both LTRs at the time of integration, but at least one point mutation in the 3' LTR has since accumulated. The possibility that these binding sites were maintained in the population due to some, currently unknown, selective advantage and biological function is should be followed up on with a larger investigation into all functional HML-2 binding sites.

Due to their possible role in pathogenicity, it is essential to study the genetic differences of HML-2 elements between individuals. Most often, this is looked at in

regards to the whole provirus by studying insertionally or allelically polymorphic proviruses and their possible contribution to disease. Thus far, however, no polymorphic provirus has been found to play a role in breast cancer tumorigenesis (109, 122). This is the first study to our knowledge to investigate the genetic differences at the single nucleotide level, by examining SNPs and indels within LTRs. Of the 63 unique binding sites that we identified, only five of them were found to be insertionally polymorphic within the 2,504 genomes mined (**Figure 5-13, Table 5-5**). All five sites were only found in the 5' LTR and therefore acquired over time. Interestingly, all but one site (IK3) was isolated from the 11p15.4 provirus. IK3 was found in the 5' LTR of 3q21.2, a human-specific provirus with an estimated age of 4.8-8.69 million years old (40).

The allele frequencies of the five polymorphic binding sites within the human population ranged from 42-52%. However, when examined by super population (East Asian, Ad Mixed American, African, European, and South Asian), four of the binding sites showed a slightly higher allele frequency (at least 60%) in Ad Mixed American and African populations (**Figure 5-14**). In general, African populations have the highest degree of genetic variance and polymorphic proviruses are known to occur at higher frequencies in individuals of African descent (41). Only one binding site, IK3, had a lower allele frequency in African populations. This binding site is present in ~50% of individuals from East Asian, Ad Mixed American, European, and South Asian populations, but is present in only ~20% of individuals from African populations. This low frequency suggests that the binding site was either unfixed at the time of human migration out of Africa or arose relatively recently in human evolution.

The RORA binding site, harbored on the 5' LTR of the 11p15.4 provirus, was the only site critical for HML-2 activation in HMLE cells that was also polymorphic. This provirus is of particular interest because it is a segmental duplication of 8p23.1c (40). 8p23.1c showed no LTR activity in either HMLE cell line, unlike 11p15.4 (**Figure 5-4**). Despite these sequences having an age estimated at about 15-28 million years old, 11p15.4 is a human-specific sequence (40). The human-specific quality of this provirus is believed to be from a segmental duplication occurring at some point after the evolutionary split between humans and chimpanzees.

After examining the RORA binding site on the 5' and 3' LTRs of both proviruses, we found the site to be identical on the 3' LTRs. The RORA site on the 5' LTRs differed by only one nucleotide, found at position 23 (**Figure 5-16**). Based on the luciferase results, this one nucleotide appears to be responsible for whether or not the binding site is active since 11p15.4 exhibited high promoter activity while 8p23.1c did not. Possessing an A at position 23 results in an active RORA binding site, as seen with 11p15.4, whereas a T in that position results in an inactive binding site, as seen with 8p23.1c. Analysis using data from the 1000 Genomes Project showed that this nucleotide establishes the polymorphism of the site; 51% of the population has an A whereas 49% of the population has a T. These results suggest that half of the human population has evolved away from the ancestral (8p23.1c) version of the LTR, towards a more active (11p15.4) version.

6.2 Future Directions

Our lab has produced an RNA-Seq protocol that is optimized for the capture of HML-2 expression and successful at identifying transcription from non-reference HML-2 elements as well as proviruses of high sequence similarity (110). The foundation of this protocol is based on Illumina MiSeq sequencing. This platform was chosen because, at the time we conducted our sequencing experiments, Illumina had a much lower error rate than some other systems, such as PacBio or Ion Torrent. Additionally, at the time, MiSeq was the only benchtop or production-scale Illumina sequencer to offer paired-end reads over 150 bp in length (259). The need for a low error rate and longer read length were crucial for proper detection of HML-2 elements. This requirement was mirrored in our results, which showed that the majority of multi-mapped reads were short, averaging 112 bp in length, and aligned to the LTR regions of proviruses with solo LTRs of highly similar sequence. However, increased read length comes at the expense of lower sequencing coverage, an issue for detecting possible novel integrations or proviruses expressed at low levels.

Recent advancements are being made in NGS technology and new Illumina HiSeq 2500 platforms can support paired-end reads with lengths up to 250 bp. This produces up to 300 million reads per single flow cell, an obvious advancement in read coverage as compared to MiSeq, which produces up to 30 million paired-end reads. Hopefully, these innovations keep advancing and future studies can be conducted using longer read lengths without compromising coverage. This combination of long reads and high coverage would allow for *de novo* assembly of proviruses, a technique that may be the best option for detecting novel, non-reference integrations.

Overall, our sequencing study was to ascertain whether our sequencing protocol could capture HML-2 expression during HMEC transformation and whether influences on host gene transcription, such as the donation of alternative promoters, splice sites, and termination signals, could be detected. We believe our methodology was successful and the next steps would be to repeat this study with a larger collection of human samples. NGS analysis of the HML-2 transcriptome in a large number of matched tumor and non-diseased samples would provide more definitive results in determining whether HML-2 expression is contributing to tumorigenesis. Although no significantly expressed provirus in our study is believed to have a viable ORF for any viral gene, protein production in these cell lines as well as any tumor sample used in future investigations, should be examined. HML-2 protein production may have use as a molecular biomarker or as a target for immunotherapy, but more samples need to be tested in order to understand the scope of their expression.

Aside from tumor samples, it would be interesting to repeat these experiments in non-diseased human tissues that are known to expression HERV-K (HML-2) at high levels, such as the placenta or embryonic stem cells. Investigating the HML-2 transcriptome in these two cell types may allow for a better understanding of the regulation of ERV expression under non-diseased conditions.

Our investigations into polymorphic binding sites highlight the importance of studying LTR SNPs and not just polymorphic integrations of whole proviruses. LTR SNPs that occur in U3 binding sites may play important roles in the differential expression of ERVs that is often seen amongst different tissue types as well as different individuals. All but one of the unique polymorphic binding sites that we found had higher

allele frequencies in African populations. In general, these populations tend to have higher levels of overall genetic variance, but are unfortunately under-represented in most genetic studies and clinical trials and therefore the implications of high genetic variance are currently unclear.

The possible role of pioneer factors in driving ERV transcription is intriguing and should be studied in more detail. We only examined unique binding sites in this analysis, but it is possible that pioneer factors may bind to more ubiquitous LTR sites. The two most well studied pioneer factors are FOXA and GATA proteins (249, 260). Interestingly, of the nine 5' LTRs that we performed transcription factor binding site analysis on, all but one provirus (7p22.1b) contained forkhead family binding sites utilized by FOXA proteins and all but two proviruses (7p22.1b and 8p23.1c) contained GATA protein binding sites. It is possible that these factors played a role in the biology of the ancestral virus, allowing for the bypass of canonical epigenetic silencing factors such as CpG methylation and promoting transcription of hypermethylated sequences.

Lastly, our investigation only looked at full-length proviruses. Solo LTRs are present in the human genome at a rate 10-fold higher than proviruses. These elements are also much smaller, with the average solo LTR being 1 kb and the average full-length provirus being 10 kb. As mentioned above, advancements in NGS technology that result in longer read lengths and better sequencing coverage are critical in order to examine solo LTR expression without substantial multi-mapping issues. Solo LTR analysis is importance since they are islands of regulatory sequences scattered throughout the genome that may be directly influencing gene transcription or aiding in the production of non-coding RNAs.

6.3 Concluding Remarks

The goal of this study was to characterize the HML-2 transcriptome and mechanisms of transcription during human mammary epithelial cell transformation. Overall, our results indicate the importance of studying ERV activty in a disease context. Although we did not find any indication of active 5' LTR use influencing nearby host gene transcription, we did uncover the donation of a 3' LTR polyadenylation site for lncRNA production at the 1q22 locus. This lncRNA, although previously annotated and known to be upregulated in breast tissue, currently has an unknown function. These observations highlight the importance of studying the donation of ERV regulatory sites in human tissue samples as they may have implications on host gene regulation and disease progression.

Our investigation into LTR *cis*-regulatory sequences also gives credence for the necessity of researching polymorphic binding sites, as they may be responsible for differential ERV expression. We found one binding site, RORA, located on the 5' LTR of the 11p15.4 provirus, that appears to be critical for the upregulation of LTR activity in HMLE cells. This binding site is only present in 51% of the human population and, importantly, appears to be evolving away from its ancestral, non-active version.

This study also demonstrates the need to investigate pioneer factors, as they may be responsible for ERV expression despite silencing attempts via CpG methylation. One site that we found, HOX-PBX, located on the 5' LTR of 3q12.3, has a 4 bp duplication that creates the human-specific binding site and PBX1, a known pioneer factor, may use this site for transcriptional activation. Future investigations will need to be conducted to determine the relationship between pioneer factors and ERV transcription. Overall,

despite these studies contributing to the knowledge of HML-2 regulation and expression, the biological significance of their activation remains unclear. Likely, their expression is merely a consequence of the transcriptional environment of a malignant cell. However, the possibility of ERV sequences being activated post-transformation and contributing to the tumorigenic process cannot be ruled out.

Bibliography

1. **Vogt PK.** 1997. Historical Introduction to the General Properties of Retroviruses. *In* Coffin JM, Hughes SH, Varmus HE (ed), Retroviruses, Cold Spring Harbor (NY).
2. **Bannert N, Kurth R.** 2006. The evolutionary dynamics of human endogenous retroviral families. Annu Rev Genomics Hum Genet **7:**149-173.
3. **Jern P, Coffin JM.** 2008. Effects of retroviruses on host genome function. Annu Rev Genet **42:**709-732.
4. **Rosenberg N.** 2011. Overview of Retrovirology. *In* Dudley J (ed), Retroviruses and insights into cancer. Springer, NY.
5. **Ryu W.** 2017. Molecular Virology of Human Pathogenic Viruses. Academic Press.
6. **Rambaut A, Posada D, Crandall KA, Holmes EC.** 2004. The causes and consequences of HIV evolution. Nat Rev Genet **5:**52-61.
7. **Fauci AS, Desrosiers RC.** 1997. Pathogenesis of HIV and SIV. *In* Coffin JM, Hughes SH, Varmus HE (ed), Retroviruses, Cold Spring Harbor (NY).
8. **Telesnitsky A, Goff SP.** 1997. Reverse Transcriptase and the Generation of Retroviral DNA. *In* Coffin JM, Hughes SH, Varmus HE (ed), Retroviruses, Cold Spring Harbor (NY).
9. **Swanstrom R, Wills JW.** 1997. Synthesis, Assembly, and Processing of Viral Proteins. *In* Coffin JM, Hughes SH, Varmus HE (ed), Retroviruses, Cold Spring Harbor (NY).
10. **Vogt VM.** 1997. Retroviral Virions and Genomes. *In* Coffin JM, Hughes SH, Varmus HE (ed), Retroviruses, Cold Spring Harbor (NY).
11. **Rabson AB, Graves BJ.** 1997. Synthesis and Processing of Viral RNA. *In* Coffin JM, Hughes SH, Varmus HE (ed), Retroviruses, Cold Spring Harbor (NY).
12. **Coffin JM, Hughes SH, Varmus HE.** 1997. The Interactions of Retroviruses and their Hosts. *In* Coffin JM, Hughes SH, Varmus HE (ed), Retroviruses, Cold Spring Harbor (NY).
13. **Goepfert PA, Shaw KL, Ritter GD, Jr., Mulligan MJ.** 1997. A sorting motif localizes the foamy virus glycoprotein to the endoplasmic reticulum. J Virol **71:**778-784.
14. **He F, Blair WS, Fukushima J, Cullen BR.** 1996. The human foamy virus Bel-1 transcription factor is a sequence-specific DNA binding protein. J Virol **70:**3902-3908.
15. **Moebes A, Enssle J, Bieniasz PD, Heinkelein M, Lindemann D, Bock M, McClure MO, Rethwilm A.** 1997. Human foamy virus reverse transcription that occurs late in the viral replication cycle. J Virol **71:**7305-7311.
16. **Rous P.** 1911. A Sarcoma of the Fowl Transmissible by an Agent Separable from the Tumor Cells. J Exp Med **13:**397-411.
17. **Ajiro M, Zheng ZM.** 2014. Oncogenes and RNA splicing of human tumor viruses. Emerg Microbes Infect **3:**e63.
18. **Rosenberg N, Jolicoeur P.** 1997. Retroviral Pathogenesis. *In* Coffin JM, Hughes SH, Varmus HE (ed), Retroviruses, Cold Spring Harbor (NY).
19. **Lee EY, Muller WJ.** 2010. Oncogenes and tumor suppressor genes. Cold Spring Harb Perspect Biol **2:**a003236.

20. **Bittner JJ.** 1936. Some Possible Effects of Nursing on the Mammary Gland Tumor Incidence in Mice. Science **84:**162.
21. **Verwoerd DW, Payne AL, York DF, Myer MS.** 1983. Isolation and preliminary characterization of the jaagsiekte retrovirus (JSRV). Onderstepoort J Vet Res **50:**309-316.
22. **Martineau D, Bowser PR, Renshaw RR, Casey JW.** 1992. Molecular characterization of a unique retrovirus associated with a fish tumor. J Virol **66:**596-599.
23. **Denner J.** 2016. Transspecies Transmission of Gammaretroviruses and the Origin of the Gibbon Ape Leukaemia Virus (GaLV) and the Koala Retrovirus (KoRV). Viruses **8**.
24. **Damania B.** 2012. Viral-Encoded Genes and Cancer. *In* Robertson ES (ed), Cancer Associated Viruses. Springer.
25. **Zhu K, Liu Q, Zhou Y, Tao C, Zhao Z, Sun J, Xu H.** 2015. Oncogenes and tumor suppressor genes: comparative genomics and network perspectives. BMC Genomics **16 Suppl 7:**S8.
26. **Wallace NA, Galloway DA.** 2016. Viral Oncogenesis: Infections that Can Lead to Cancer. *In* Katze MG, Korth MJ, Law GL, Nathanson N (ed), Viral Pathogenesis: From Basics to Systems Biology. Academic Press.
27. **Beemon K, Rosenberg N.** 2012. Mechanisms of Oncogenesis by Avian and Murine Retroviruses. *In* Robertson ES (ed), Cancer Associated Viruses. Springer.
28. **Beemon KL, Bolisetty M.** 2011. Mechanisms of oncogenesis by retroviruses. *In* Dudley J (ed), Retroviruses and insights into cancer. Springer, NY.
29. **Herman SA, Coffin JM.** 1986. Differential transcription from the long terminal repeats of integrated avian leukosis virus DNA. J Virol **60:**497-505.
30. **Payne GS, Bishop JM, Varmus HE.** 1982. Multiple arrangements of viral DNA and an activated host oncogene in bursal lymphomas. Nature **295:**209-214.
31. **Li JP, Baltimore D.** 1991. Mechanism of leukemogenesis induced by mink cell focus-forming murine leukemia viruses. J Virol **65:**2408-2414.
32. **Zhang J, Randall MS, Loyd MR, Li W, Schweers RL, Persons DA, Rehg JE, Noguchi CT, Ihle JN, Ney PA.** 2006. Role of erythropoietin receptor signaling in Friend virus-induced erythroblastosis and polycythemia. Blood **107:**73-78.
33. **Hoatlin ME, Kozak SL, Lilly F, Chakraborti A, Kozak CA, Kabat D.** 1990. Activation of erythropoietin receptors by Friend viral gp55 and by erythropoietin and down-modulation by the murine Fv-2r resistance gene. Proc Natl Acad Sci U S A **87:**9985-9989.
34. **Yugawa T, Amanuma H.** 1998. Sequence flexibility in the polytropic env gp70-derived region of the membrane glycoprotein (gp55) of Friend spleen focus-forming virus affects its biological activity. J Virol **72:**2272-2279.
35. **Holt MP, Shevach EM, Punkosdy GA.** 2013. Endogenous mouse mammary tumor viruses (mtv): new roles for an old virus in cancer, infection, and immunity. Front Oncol **3:**287.
36. **Acha-Orbea H, MacDonald HR.** 1995. Superantigens of mouse mammary tumor virus. Annu Rev Immunol **13:**459-486.
37. **Johnson WE.** 2015. Endogenous Retroviruses in the Genomics Era. Annu Rev Virol **2:**135-159.

38. **Lower R, Lower J, Tondera-Koch C, Kurth R.** 1993. A general method for the identification of transcribed retrovirus sequences (R-U5 PCR) reveals the expression of the human endogenous retrovirus loci HERV-H and HERV-K in teratocarcinoma cells. Virology **192:**501-511.
39. **Babaian A, Mager DL.** 2016. Endogenous retroviral promoter exaptation in human cancer. Mob DNA **7:**24.
40. **Subramanian RP, Wildschutte JH, Russo C, Coffin JM.** 2011. Identification, characterization, and comparative genomic distribution of the HERV-K (HML-2) group of human endogenous retroviruses. Retrovirology **8:**90.
41. **Wildschutte JH, Williams ZH, Montesion M, Subramanian RP, Kidd JM, Coffin JM.** 2016. Discovery of unfixed endogenous retrovirus insertions in diverse human populations. Proc Natl Acad Sci U S A doi:10.1073/pnas.1602336113.
42. **Perot P, Bolze P, Mallet F.** 2012. From Viruses to Genes: Syncytins. *In* Witzany G (ed), Viruses: Essential Agents of Life. Springer.
43. **Lower R, Lower J, Kurth R.** 1996. The viruses in all of us: characteristics and biological significance of human endogenous retrovirus sequences. Proc Natl Acad Sci U S A **93:**5177-5184.
44. **Xu W, Eiden MV.** 2015. Koala Retroviruses: Evolution and Disease Dynamics. Annu Rev Virol **2:**119-134.
45. **Armezzani A, Murphy L, Spencer TE, Palmarini M, Arnaud F.** 2012. The Evolutionary Interplay Between Exogenous and Endogenous Sheep Betaretroviruses. *In* Witzany G (ed), Viruses: Essential Agents of Life. Springer.
46. **Eiden MV, Taliaferro DL.** 2011. Emerging Retroviruses and Cancer. *In* Dudley J (ed), Retroviruses and Insights into Cancer. Springer.
47. **Tarlinton RE.** 2012. Koala Retrovirus Endogenization in Action. *In* Witzany G (ed), Viruses: Essential Agents of Life. Springer.
48. **Naville M, Warren IA, Haftek-Terreau Z, Chalopin D, Brunet F, Levin P, Galiana D, Volff JN.** 2016. Not so bad after all: retroviruses and long terminal repeat retrotransposons as a source of new genes in vertebrates. Clin Microbiol Infect **22:**312-323.
49. **Ruprecht K, Mayer J, Sauter M, Roemer K, Mueller-Lantzsch N.** 2008. Endogenous retroviruses and cancer. Cell Mol Life Sci **65:**3366-3382.
50. **Wang-Johanning F, Frost AR, Johanning GL, Khazaeli MB, LoBuglio AF, Shaw DR, Strong TV.** 2001. Expression of human endogenous retrovirus k envelope transcripts in human breast cancer. Clin Cancer Res **7:**1553-1560.
51. **Dudley JP, Mertz JA, Bhadra S, Palmarini M, Kozak CA.** 2011. Endogenous Retroviruses and Cancer. *In* Dudley J (ed), Retroviruses and Insights into Cancer. Springer.
52. **van de Lagemaat LN, Medstrand P, Mager DL.** 2006. Multiple effects govern endogenous retrovirus survival patterns in human gene introns. Genome Biol **7:**R86.
53. **Ruda VM, Akopov SB, Trubetskoy DO, Manuylov NL, Vetchinova AS, Zavalova LL, Nikolaev LG, Sverdlov ED.** 2004. Tissue specificity of enhancer and promoter activities of a HERV-K(HML-2) LTR. Virus Res **104:**11-16.

54. **Young GR, Stoye JP, Kassiotis G.** 2013. Are human endogenous retroviruses pathogenic? An approach to testing the hypothesis. Bioessays **35:**794-803.
55. **Wang J, Xie G, Singh M, Ghanbarian AT, Rasko T, Szvetnik A, Cai H, Besser D, Prigione A, Fuchs NV, Schumann GG, Chen W, Lorincz MC, Ivics Z, Hurst LD, Izsvak Z.** 2014. Primate-specific endogenous retrovirus-driven transcription defines naive-like stem cells. Nature **516:**405-409.
56. **Lu X, Sachs F, Ramsay L, Jacques PE, Goke J, Bourque G, Ng HH.** 2014. The retrovirus HERVH is a long noncoding RNA required for human embryonic stem cell identity. Nat Struct Mol Biol **21:**423-425.
57. **Samuelson LC, Wiebauer K, Snow CM, Meisler MH.** 1990. Retroviral and pseudogene insertion sites reveal the lineage of human salivary and pancreatic amylase genes from a single gene during primate evolution. Mol Cell Biol **10:**2513-2520.
58. **Ting CN, Rosenberg MP, Snow CM, Samuelson LC, Meisler MH.** 1992. Endogenous retroviral sequences are required for tissue-specific expression of a human salivary amylase gene. Genes Dev **6:**1457-1465.
59. **Kapitonov VV, Jurka J.** 1999. The long terminal repeat of an endogenous retrovirus induces alternative splicing and encodes an additional carboxy-terminal sequence in the human leptin receptor. J Mol Evol **48:**248-251.
60. **True JR, Carroll SB.** 2002. Gene co-option in physiological and morphological evolution. Annu Rev Cell Dev Biol **18:**53-80.
61. **Sanetra M, Begemann G, Becker MB, Meyer A.** 2005. Conservation and co-option in developmental programmes: the importance of homology relationships. Front Zool **2:**15.
62. **Di Cristofano A, Strazzullo M, Longo L, La Mantia G.** 1995. Characterization and genomic mapping of the ZNF80 locus: expression of this zinc-finger gene is driven by a solitary LTR of ERV9 endogenous retroviral family. Nucleic Acids Res **23:**2823-2830.
63. **Medstrand P, Landry JR, Mager DL.** 2001. Long terminal repeats are used as alternative promoters for the endothelin B receptor and apolipoprotein C-I genes in humans. J Biol Chem **276:**1896-1903.
64. **Dunn CA, Romanish MT, Gutierrez LE, van de Lagemaat LN, Mager DL.** 2006. Transcription of two human genes from a bidirectional endogenous retrovirus promoter. Gene **366:**335-342.
65. **Feuchter-Murthy AE, Freeman JD, Mager DL.** 1993. Splicing of a human endogenous retrovirus to a novel phospholipase A2 related gene. Nucleic Acids Res **21:**135-143.
66. **Romanish MT, Lock WM, van de Lagemaat LN, Dunn CA, Mager DL.** 2007. Repeated recruitment of LTR retrotransposons as promoters by the anti-apoptotic locus NAIP during mammalian evolution. PLoS Genet **3:**e10.
67. **Kato N, Shimotohno K, VanLeeuwen D, Cohen M.** 1990. Human proviral mRNAs down regulated in choriocarcinoma encode a zinc finger protein related to Kruppel. Mol Cell Biol **10:**4401-4405.
68. **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 beta-globin locus control region. Genomics **54:**542-555.

69. **Schulte AM, Lai S, Kurtz A, Czubayko F, Riegel AT, Wellstein A.** 1996. Human trophoblast and choriocarcinoma expression of the growth factor pleiotrophin attributable to germ-line insertion of an endogenous retrovirus. Proc Natl Acad Sci U S A **93:**14759-14764.
70. **Soygur B, Sati L.** 2016. The role of syncytins in human reproduction and reproductive organ cancers. Reproduction **152:**R167-178.
71. **Dupressoir A, Vernochet C, Harper F, Guegan J, Dessen P, Pierron G, Heidmann T.** 2011. A pair of co-opted retroviral envelope syncytin genes is required for formation of the two-layered murine placental syncytiotrophoblast. Proc Natl Acad Sci U S A **108:**E1164-1173.
72. **Lavialle C, Cornelis G, Dupressoir A, Esnault C, Heidmann O, Vernochet C, Heidmann T.** 2013. Paleovirology of 'syncytins', retroviral env genes exapted for a role in placentation. Philos Trans R Soc Lond B Biol Sci **368:**20120507.
73. **Malik HS.** 2012. Retroviruses push the envelope for mammalian placentation. Proc Natl Acad Sci U S A **109:**2184-2185.
74. **Blond JL, Lavillette D, Cheynet V, Bouton O, Oriol G, Chapel-Fernandes S, Mandrand B, Mallet F, Cosset FL.** 2000. An envelope glycoprotein of the human endogenous retrovirus HERV-W is expressed in the human placenta and fuses cells expressing the type D mammalian retrovirus receptor. J Virol **74:**3321-3329.
75. **Heidmann O, Vernochet C, Dupressoir A, Heidmann T.** 2009. Identification of an endogenous retroviral envelope gene with fusogenic activity and placenta-specific expression in the rabbit: a new "syncytin" in a third order of mammals. Retrovirology **6:**107.
76. **Dupressoir A, Marceau G, Vernochet C, Benit L, Kanellopoulos C, Sapin V, Heidmann T.** 2005. Syncytin-A and syncytin-B, two fusogenic placenta-specific murine envelope genes of retroviral origin conserved in Muridae. Proc Natl Acad Sci U S A **102:**725-730.
77. **Mi S, Lee X, Li X, Veldman GM, Finnerty H, Racie L, LaVallie E, Tang XY, Edouard P, Howes S, Keith JC, Jr., McCoy JM.** 2000. Syncytin is a captive retroviral envelope protein involved in human placental morphogenesis. Nature **403:**785-789.
78. **Renard M, Varela PF, Letzelter C, Duquerroy S, Rey FA, Heidmann T.** 2005. Crystal structure of a pivotal domain of human syncytin-2, a 40 million years old endogenous retrovirus fusogenic envelope gene captured by primates. J Mol Biol **352:**1029-1034.
79. **Vernochet C, Heidmann O, Dupressoir A, Cornelis G, Dessen P, Catzeflis F, Heidmann T.** 2011. A syncytin-like endogenous retrovirus envelope gene of the guinea pig specifically expressed in the placenta junctional zone and conserved in Caviomorpha. Placenta **32:**885-892.
80. **Grow EJ, Flynn RA, Chavez SL, Bayless NL, Wossidlo M, Wesche DJ, Martin L, Ware CB, Blish CA, Chang HY, Reijo Pera RA, Wysocka J.** 2015. Intrinsic retroviral reactivation in human preimplantation embryos and pluripotent cells. Nature doi:10.1038/nature14308.

81. **Yap MW, Colbeck E, Ellis SA, Stoye JP.** 2014. Evolution of the retroviral restriction gene Fv1: inhibition of non-MLV retroviruses. PLoS Pathog **10:**e1003968.
82. **Li W, Yap MW, Voss V, Stoye JP.** 2016. Expression levels of Fv1: effects on retroviral restriction specificities. Retrovirology **13:**42.
83. **Nethe M, Berkhout B, van der Kuyl AC.** 2005. Retroviral superinfection resistance. Retrovirology **2:**52.
84. **Robinson HL, Astrin SM, Senior AM, Salazar FH.** 1981. Host Susceptibility to endogenous viruses: defective, glycoprotein-expressing proviruses interfere with infections. J Virol **40:**745-751.
85. **Mei M, Ye J, Qin A, Wang L, Hu X, Qian K, Shao H.** 2015. Identification of novel viral receptors with cell line expressing viral receptor-binding protein. Sci Rep **5:**7935.
86. **Palmarini M, Mura M, Spencer TE.** 2004. Endogenous betaretroviruses of sheep: teaching new lessons in retroviral interference and adaptation. J Gen Virol **85:**1-13.
87. **Contreras-Galindo R, Kaplan MH, Contreras-Galindo AC, Gonzalez-Hernandez MJ, Ferlenghi I, Giusti F, Lorenzo E, Gitlin SD, Dosik MH, Yamamura Y, Markovitz DM.** 2012. Characterization of human endogenous retroviral elements in the blood of HIV-1-infected individuals. J Virol **86:**262-276.
88. **Young GR, Eksmond U, Salcedo R, Alexopoulou L, Stoye JP, Kassiotis G.** 2012. Resurrection of endogenous retroviruses in antibody-deficient mice. Nature **491:**774-778.
89. **Dewannieux M, Harper F, Richaud A, Letzelter C, Ribet D, Pierron G, Heidmann T.** 2006. Identification of an infectious progenitor for the multiple-copy HERV-K human endogenous retroelements. Genome Res **16:**1548-1556.
90. **Lee YN, Bieniasz PD.** 2007. Reconstitution of an infectious human endogenous retrovirus. PLoS Pathog **3:**e10.
91. **Gotzinger N, Sauter M, Roemer K, Mueller-Lantzsch N.** 1996. Regulation of human endogenous retrovirus-K Gag expression in teratocarcinoma cell lines and human tumours. J Gen Virol **77 (Pt 12):**2983-2990.
92. **Florl AR, Lower R, Schmitz-Drager BJ, Schulz WA.** 1999. DNA methylation and expression of LINE-1 and HERV-K provirus sequences in urothelial and renal cell carcinomas. Br J Cancer **80:**1312-1321.
93. **Saxonov S, Berg P, Brutlag DL.** 2006. A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters. Proc Natl Acad Sci U S A **103:**1412-1417.
94. **Deaton AM, Bird A.** 2011. CpG islands and the regulation of transcription. Genes Dev **25:**1010-1022.
95. **Conley AB, Jordan IK.** 2012. Endogenous Retroviruses and the Epigenome. *In* Witzany G (ed), Viruses: Essential Agents of Life. Springer.
96. **Nan X, Ng HH, Johnson CA, Laherty CD, Turner BM, Eisenman RN, Bird A.** 1998. Transcriptional repression by the methyl-CpG-binding protein MeCP2 involves a histone deacetylase complex. Nature **393:**386-389.

97. **Du Q, Luu PL, Stirzaker C, Clark SJ.** 2015. Methyl-CpG-binding domain proteins: readers of the epigenome. Epigenomics **7:**1051-1073.
98. **Reiss D, Zhang Y, Mager DL.** 2007. Widely variable endogenous retroviral methylation levels in human placenta. Nucleic Acids Res **35:**4743-4754.
99. **Moyes D, Griffiths DJ, Venables PJ.** 2007. Insertional polymorphisms: a new lease of life for endogenous retroviruses in human disease. Trends Genet **23:**326-333.
100. **Slokar G, Hasler G.** 2015. Human Endogenous Retroviruses as Pathogenic Factors in the Development of Schizophrenia. Front Psychiatry **6:**183.
101. **Herve CA, Lugli EB, Brand A, Griffiths DJ, Venables PJ.** 2002. Autoantibodies to human endogenous retrovirus-K are frequently detected in health and disease and react with multiple epitopes. Clin Exp Immunol **128:**75-82.
102. **Rolland A, Jouvin-Marche E, Saresella M, Ferrante P, Cavaretta R, Creange A, Marche P, Perron H.** 2005. Correlation between disease severity and in vitro cytokine production mediated by MSRV (multiple sclerosis associated retroviral element) envelope protein in patients with multiple sclerosis. J Neuroimmunol **160:**195-203.
103. **Green JE, Hinrichs SH, Vogel J, Jay G.** 1989. Exocrinopathy resembling Sjogren's syndrome in HTLV-1 tax transgenic mice. Nature **341:**72-74.
104. **Lower R, Boller K, Hasenmaier B, Korbmacher C, Muller-Lantzsch N, Lower J, Kurth R.** 1993. Identification of human endogenous retroviruses with complex mRNA expression and particle formation. Proc Natl Acad Sci U S A **90:**4480-4484.
105. **Wang-Johanning F, Frost AR, Jian B, Epp L, Lu DW, Johanning GL.** 2003. Quantitation of HERV-K env gene expression and splicing in human breast cancer. Oncogene **22:**1528-1535.
106. **Muster T, Waltenberger A, Grassauer A, Hirschl S, Caucig P, Romirer I, Fodinger D, Seppele H, Schanab O, Magin-Lachmann C, Lower R, Jansen B, Pehamberger H, Wolff K.** 2003. An endogenous retrovirus derived from human melanoma cells. Cancer Res **63:**8735-8741.
107. **Buscher K, Trefzer U, Hofmann M, Sterry W, Kurth R, Denner J.** 2005. Expression of human endogenous retrovirus K in melanomas and melanoma cell lines. Cancer Res **65:**4172-4180.
108. **Bieda K, Hoffmann A, Boller K.** 2001. Phenotypic heterogeneity of human endogenous retrovirus particles produced by teratocarcinoma cell lines. J Gen Virol **82:**591-596.
109. **Wildschutte JH, Ram D, Subramanian R, Stevens VL, Coffin JM.** 2014. The distribution of insertionally polymorphic endogenous retroviruses in breast cancer patients and cancer-free controls. Retrovirology **11:**62.
110. **Bhardwaj N, Montesion M, Roy F, Coffin JM.** 2015. Differential expression of HERV-K (HML-2) proviruses in cells and virions of the teratocarcinoma cell line Tera-1. Viruses **7:**939-968.
111. **King AMQ, Adams MJ, Carstens EB, Lefkowitz EJ.** 2012. Virus Taxonomy: Ninth Report of the International Committee on Taxonomy of Viruses. Elsevier.

112. **Karolchik D, Hinrichs AS, Furey TS, Roskin KM, Sugnet CW, Haussler D, Kent WJ.** 2004. The UCSC Table Browser data retrieval tool. Nucleic Acids Res **32:**D493-496.
113. **Seifarth W, Skladny H, Krieg-Schneider F, Reichert A, Hehlmann R, Leib-Mosch C.** 1995. Retrovirus-like particles released from the human breast cancer cell line T47-D display type B- and C-related endogenous retroviral sequences. J Virol **69:**6408-6416.
114. **Simpson GR, Patience C, Lower R, Tonjes RR, Moore HD, Weiss RA, Boyd MT.** 1996. Endogenous D-type (HERV-K) related sequences are packaged into retroviral particles in the placenta and possess open reading frames for reverse transcriptase. Virology **222:**451-456.
115. **Ruprecht K, Ferreira H, Flockerzi A, Wahl S, Sauter M, Mayer J, Mueller-Lantzsch N.** 2008. Human endogenous retrovirus family HERV-K(HML-2) RNA transcripts are selectively packaged into retroviral particles produced by the human germ cell tumor line Tera-1 and originate mainly from a provirus on chromosome 22q11.21. J Virol **82:**10008-10016.
116. **Seifarth W, Baust C, Schon U, Reichert A, Hehlmann R, Leib-Mosch C.** 2000. HERV-IP-T47D, a novel type C-related human endogenous retroviral sequence derived from T47D particles. AIDS Res Hum Retroviruses **16:**471-480.
117. **Zahn J, Kaplan MH, Fischer S, Dai M, Meng F, Saha AK, Cervantes P, Chan SM, Dube D, Omenn GS, Markovitz DM, Contreras-Galindo R.** 2015. Expansion of a novel endogenous retrovirus throughout the pericentromeres of modern humans. Genome Biol **16:**74.
118. **Belshaw R, Pereira V, Katzourakis A, Talbot G, Paces J, Burt A, Tristem M.** 2004. Long-term reinfection of the human genome by endogenous retroviruses. Proc Natl Acad Sci U S A **101:**4894-4899.
119. **Golan M, Hizi A, Resau JH, Yaal-Hahoshen N, Reichman H, Keydar I, Tsarfaty I.** 2008. Human endogenous retrovirus (HERV-K) reverse transcriptase as a breast cancer prognostic marker. Neoplasia **10:**521-533.
120. **Wang-Johanning F, Radvanyi L, Rycaj K, Plummer JB, Yan P, Sastry KJ, Piyathilake CJ, Hunt KK, Johanning GL.** 2008. Human endogenous retrovirus K triggers an antigen-specific immune response in breast cancer patients. Cancer Res **68:**5869-5877.
121. **Moslehi R, Freedman E, Zeinomar N, Veneroso C, Levine PH.** 2016. Importance of hereditary and selected environmental risk factors in the etiology of inflammatory breast cancer: a case-comparison study. BMC Cancer **16:**334.
122. **Burmeister T, Ebert AD, Pritze W, Loddenkemper C, Schwartz S, Thiel E.** 2004. Insertional polymorphisms of endogenous HERV-K113 and HERV-K115 retroviruses in breast cancer patients and age-matched controls. AIDS Res Hum Retroviruses **20:**1223-1229.
123. **Nguyen A, Yoshida M, Goodarzi H, Tavazoie SF.** 2016. Highly variable cancer subpopulations that exhibit enhanced transcriptome variability and metastatic fitness. Nat Commun **7:**11246.
124. **Vogelstein B, Kinzler KW.** 1993. The multistep nature of cancer. Trends Genet **9:**138-141.

125. **Hanahan D, Weinberg RA.** 2011. Hallmarks of cancer: the next generation. Cell **144:**646-674.
126. **Zhao J, Rycaj K, Geng S, Li M, Plummer JB, Yin B, Liu H, Xu X, Zhang Y, Yan Y, Glynn SA, Dorsey TH, Ambs S, Johanning GL, Gu L, Wang-Johanning F.** 2011. Expression of Human Endogenous Retrovirus Type K Envelope Protein is a Novel Candidate Prognostic Marker for Human Breast Cancer. Genes Cancer **2:**914-922.
127. **Weiss RA.** 2012. Oncogenic Viruses and Cancer Transmission. *In* Roberston ES (ed), Cancer Associated Viruses. Sringer.
128. **Buehring GC, Shen HM, Jensen HM, Jin DL, Hudes M, Block G.** 2015. Exposure to Bovine Leukemia Virus Is Associated with Breast Cancer: A Case-Control Study. PLoS One **10:**e0134304.
129. **Lawson JS, Glenn WK, Salyakina D, Delprado W, Clay R, Antonsson A, Heng B, Miyauchi S, Tran DD, Ngan CC, Lutze-Mann L, Whitaker NJ.** 2015. Human Papilloma Viruses and Breast Cancer. Front Oncol **5:**277.
130. **Ono M, Kawakami M, Ushikubo H.** 1987. Stimulation of expression of the human endogenous retrovirus genome by female steroid hormones in human breast cancer cell line T47D. J Virol **61:**2059-2062.
131. **Wang-Johanning F, Liu J, Rycaj K, Huang M, Tsai K, Rosen DG, Chen DT, Lu DW, Barnhart KF, Johanning GL.** 2007. Expression of multiple human endogenous retrovirus surface envelope proteins in ovarian cancer. Int J Cancer **120:**81-90.
132. **Iramaneerat K, Rattanatunyong P, Khemapech N, Triratanachat S, Mutirangura A.** 2011. HERV-K hypomethylation in ovarian clear cell carcinoma is associated with a poor prognosis and platinum resistance. Int J Gynecol Cancer **21:**51-57.
133. **Schiavetti F, Thonnard J, Colau D, Boon T, Coulie PG.** 2002. A human endogenous retroviral sequence encoding an antigen recognized on melanoma by cytolytic T lymphocytes. Cancer Res **62:**5510-5516.
134. **Serafino A, Balestrieri E, Pierimarchi P, Matteucci C, Moroni G, Oricchio E, Rasi G, Mastino A, Spadafora C, Garaci E, Vallebona PS.** 2009. The activation of human endogenous retrovirus K (HERV-K) is implicated in melanoma cell malignant transformation. Exp Cell Res **315:**849-862.
135. **Depil S, Roche C, Dussart P, Prin L.** 2002. Expression of a human endogenous retrovirus, HERV-K, in the blood cells of leukemia patients. Leukemia **16:**254-259.
136. **Contreras-Galindo R, Kaplan MH, Leissner P, Verjat T, Ferlenghi I, Bagnoli F, Giusti F, Dosik MH, Hayes DF, Gitlin SD, Markovitz DM.** 2008. Human endogenous retrovirus K (HML-2) elements in the plasma of people with lymphoma and breast cancer. J Virol **82:**9329-9336.
137. **Wang-Johanning F, Li M, Esteva FJ, Hess KR, Yin B, Rycaj K, Plummer JB, Garza JG, Ambs S, Johanning GL.** 2014. Human endogenous retrovirus type K antibodies and mRNA as serum biomarkers of early-stage breast cancer. Int J Cancer **134:**587-595.
138. **Wang-Johanning F, Rycaj K, Plummer JB, Li M, Yin B, Frerich K, Garza JG, Shen J, Lin K, Yan P, Glynn SA, Dorsey TH, Hunt KK, Ambs S,**

Johanning GL. 2012. Immunotherapeutic potential of anti-human endogenous retrovirus-K envelope protein antibodies in targeting breast tumors. J Natl Cancer Inst **104:**189-210.

139. **Habraken V, van Nijnatten TJ, de Munck L, Moossdorff M, Heuts EM, Lobbes MB, Smidt ML.** 2017. Does the TNM classification of solitary internal mammary lymph node metastases in breast cancer still apply? Breast Cancer Res Treat **161:**483-489.
140. **Kao J, Salari K, Bocanegra M, Choi YL, Girard L, Gandhi J, Kwei KA, Hernandez-Boussard T, Wang P, Gazdar AF, Minna JD, Pollack JR.** 2009. Molecular profiling of breast cancer cell lines defines relevant tumor models and provides a resource for cancer gene discovery. PLoS One **4:**e6146.
141. **Holliday DL, Speirs V.** 2011. Choosing the right cell line for breast cancer research. Breast Cancer Res **13:**215.
142. **Ellingjord-Dale M, Vos L, Tretli S, Hofvind S, Dos-Santos-Silva I, Ursin G.** 2017. Parity, hormones and breast cancer subtypes - results from a large nested case-control study in a national screening program. Breast Cancer Res **19:**10.
143. **Reynier F, Verjat T, Turrel F, Imbert PE, Marotte H, Mougin B, Miossec P.** 2009. Increase in human endogenous retrovirus HERV-K (HML-2) viral load in active rheumatoid arthritis. Scand J Immunol **70:**295-299.
144. **Lower R, Tonjes RR, Korbmacher C, Kurth R, Lower J.** 1995. Identification of a Rev-related protein by analysis of spliced transcripts of the human endogenous retroviruses HTDV/HERV-K. J Virol **69:**141-149.
145. **Boese A, Galli U, Geyer M, Sauter M, Mueller-Lantzsch N.** 2001. The Rev/Rex homolog HERV-K cORF multimerizes via a C-terminal domain. FEBS Lett **493:**117-121.
146. **Hanke K, Chudak C, Kurth R, Bannert N.** 2013. The Rec protein of HERV-K(HML-2) upregulates androgen receptor activity by binding to the human small glutamine-rich tetratricopeptide repeat protein (hSGT). Int J Cancer **132:**556-567.
147. **Armbruester V, Sauter M, Krautkraemer E, Meese E, Kleiman A, Best B, Roemer K, Mueller-Lantzsch N.** 2002. A novel gene from the human endogenous retrovirus K expressed in transformed cells. Clin Cancer Res **8:**1800-1807.
148. **Buscher K, Hahn S, Hofmann M, Trefzer U, Ozel M, Sterry W, Lower J, Lower R, Kurth R, Denner J.** 2006. Expression of the human endogenous retrovirus-K transmembrane envelope, Rec and Np9 proteins in melanomas and melanoma cell lines. Melanoma Res **16:**223-234.
149. **Contreras-Galindo R, Lopez P, Velez R, Yamamura Y.** 2007. HIV-1 infection increases the expression of human endogenous retroviruses type K (HERV-K) in vitro. AIDS Res Hum Retroviruses **23:**116-122.
150. **Gross H, Barth S, Pfuhl T, Willnecker V, Spurk A, Gurtsevitch V, Sauter M, Hu B, Noessner E, Mueller-Lantzsch N, Kremmer E, Grasser FA.** 2011. The NP9 protein encoded by the human endogenous retrovirus HERV-K(HML-2) negatively regulates gene activation of the Epstein-Barr virus nuclear antigen 2 (EBNA2). Int J Cancer **129:**1105-1115.

151. **Stengel S, Fiebig U, Kurth R, Denner J.** 2010. Regulation of human endogenous retrovirus-K expression in melanomas by CpG methylation. Genes Chromosomes Cancer **49:**401-411.
152. **Rhyu DW, Kang YJ, Ock MS, Eo JW, Choi YH, Kim WJ, Leem SH, Yi JM, Kim HS, Cha HJ.** 2014. Expression of human endogenous retrovirus env genes in the blood of breast cancer patients. Int J Mol Sci **15:**9173-9183.
153. **Stauffer Y, Theiler G, Sperisen P, Lebedev Y, Jongeneel CV.** 2004. Digital expression profiles of human endogenous retroviral families in normal and cancerous tissues. Cancer Immun **4:**2.
154. **Krishnamurthy J, Rabinovich BA, Mi T, Switzer KC, Olivares S, Maiti SN, Plummer JB, Singh H, Kumaresan PR, Huls HM, Wang-Johanning F, Cooper LJ.** 2015. Genetic Engineering of T Cells to Target HERV-K, an Ancient Retrovirus on Melanoma. Clin Cancer Res **21:**3241-3251.
155. **Zhou F, Krishnamurthy J, Wei Y, Li M, Hunt K, Johanning GL, Cooper LJ, Wang-Johanning F.** 2015. Chimeric antigen receptor T cells targeting HERV-K inhibit breast cancer and its metastasis through downregulation of Ras. Oncoimmunology **4:**e1047582.
156. **Gonzalez-Hernandez MJ, Cavalcoli JD, Sartor MA, Contreras-Galindo R, Meng F, Dai M, Dube D, Saha AK, Gitlin SD, Omenn GS, Kaplan MH, Markovitz DM.** 2014. Regulation of the human endogenous retrovirus K (HML-2) transcriptome by the HIV-1 Tat protein. J Virol **88:**8924-8935.
157. **Sheehy AM, Gaddis NC, Choi JD, Malim MH.** 2002. Isolation of a human gene that inhibits HIV-1 infection and is suppressed by the viral Vif protein. Nature **418:**646-650.
158. **Katayama S, Tomaru Y, Kasukawa T, Waki K, Nakanishi M, Nakamura M, Nishida H, Yap CC, Suzuki M, Kawai J, Suzuki H, Carninci P, Hayashizaki Y, Wells C, Frith M, Ravasi T, Pang KC, Hallinan J, Mattick J, Hume DA, Lipovich L, Batalov S, Engstrom PG, Mizuno Y, Faghihi MA, Sandelin A, Chalk AM, Mottagui-Tabar S, Liang Z, Lenhard B, Wahlestedt C, Group RGER, Genome Science G, Consortium F.** 2005. Antisense transcription in the mammalian transcriptome. Science **309:**1564-1566.
159. **Wilson RC, Doudna JA.** 2013. Molecular mechanisms of RNA interference. Annu Rev Biophys **42:**217-239.
160. **Butler JE, Kadonaga JT.** 2002. The RNA polymerase II core promoter: a key component in the regulation of gene expression. Genes Dev **16:**2583-2592.
161. **He Y, Fang J, Taatjes DJ, Nogales E.** 2013. Structural visualization of key steps in human transcription initiation. Nature **495:**481-486.
162. **Fuchs NV, Kraft M, Tondera C, Hanschmann KM, Lower J, Lower R.** 2011. Expression of the human endogenous retrovirus (HERV) group HML-2/HERV-K does not depend on canonical promoter elements but is regulated by transcription factors Sp1 and Sp3. J Virol **85:**3436-3448.
163. **Katoh I, Mirova A, Kurata S, Murakami Y, Horikawa K, Nakakuki N, Sakai T, Hashimoto K, Maruyama A, Yonaga T, Fukunishi N, Moriishi K, Hirai H.** 2011. Activation of the long terminal repeat of human endogenous retrovirus K by melanoma-specific transcription factor MITF-M. Neoplasia **13:**1081-1092.

164. **Goering W, Schmitt K, Dostert M, Schaal H, Deenen R, Mayer J, Schulz WA.** 2015. Human endogenous retrovirus HERV-K(HML-2) activity in prostate cancer is dominated by a few loci. Prostate doi:10.1002/pros.23095.
165. **Vaquerizas JM, Kummerfeld SK, Teichmann SA, Luscombe NM.** 2009. A census of human transcription factors: function, expression and evolution. Nat Rev Genet **10:**252-263.
166. **Lee BK, Bhinge AA, Battenhouse A, McDaniell RM, Liu Z, Song L, Ni Y, Birney E, Lieb JD, Furey TS, Crawford GE, Iyer VR.** 2012. Cell-type specific and combinatorial usage of diverse transcription factors revealed by genome-wide binding studies in multiple human cells. Genome Res **22:**9-24.
167. **Reuter S, Gupta SC, Chaturvedi MM, Aggarwal BB.** 2010. Oxidative stress, inflammation, and cancer: how are they linked? Free Radic Biol Med **49:**1603-1616.
168. **Teschendorff AE, Zheng SC, Feber A, Yang Z, Beck S, Widschwendter M.** 2016. The multi-omic landscape of transcription factor inactivation in cancer. Genome Med **8:**89.
169. **Schmitt K, Reichrath J, Roesch A, Meese E, Mayer J.** 2013. Transcriptional profiling of human endogenous retrovirus group HERV-K(HML-2) loci in melanoma. Genome Biol Evol **5:**307-328.
170. **Knossl M, Lower R, Lower J.** 1999. Expression of the human endogenous retrovirus HTDV/HERV-K is enhanced by cellular transcription factor YY1. J Virol **73:**1254-1261.
171. **Kreimer U, Schulz WA, Koch A, Niegisch G, Goering W.** 2013. HERV-K and LINE-1 DNA Methylation and Reexpression in Urothelial Carcinoma. Front Oncol **3:**255.
172. **Lavie L, Kitova M, Maldener E, Meese E, Mayer J.** 2005. CpG methylation directly regulates transcriptional activity of the human endogenous retrovirus family HERV-K(HML-2). J Virol **79:**876-883.
173. **Levy S, Sutton G, Ng PC, Feuk L, Halpern AL, Walenz BP, Axelrod N, Huang J, Kirkness EF, Denisov G, Lin Y, MacDonald JR, Pang AW, Shago M, Stockwell TB, Tsiamouri A, Bafna V, Bansal V, Kravitz SA, Busam DA, Beeson KY, McIntosh TC, Remington KA, Abril JF, Gill J, Borman J, Rogers YH, Frazier ME, Scherer SW, Strausberg RL, Venter JC.** 2007. The diploid genome sequence of an individual human. PLoS Biol **5:**e254.
174. **Henn BM, Cavalli-Sforza LL, Feldman MW.** 2012. The great human expansion. Proc Natl Acad Sci U S A **109:**17758-17764.
175. **Campbell MC, Tishkoff SA.** 2008. African genetic diversity: implications for human demographic history, modern human origins, and complex disease mapping. Annu Rev Genomics Hum Genet **9:**403-433.
176. **Macfarlane C, Simmonds P.** 2004. Allelic variation of HERV-K(HML-2) endogenous retroviral elements in human populations. J Mol Evol **59:**642-656.
177. **Chandra Gupta S, Nandan Tripathi Y.** 2016. Potential of long non-coding RNAs in cancer patients: From biomarkers to therapeutic targets. Int J Cancer doi:10.1002/ijc.30546.

178. **Podnar J, Deiderick H, Huerta G, Hunicke-Smith S.** 2014. Next-Generation Sequencing RNA-Seq Library Construction. Curr Protoc Mol Biol **106:**4 21 21-19.
179. **Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, Pimentel H, Salzberg SL, Rinn JL, Pachter L.** 2012. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nat Protoc **7:**562-578.
180. **Schmitt K, Richter C, Backes C, Meese E, Ruprecht K, Mayer J.** 2013. Comprehensive analysis of human endogenous retrovirus group HERV-W locus transcription in multiple sclerosis brain lesions by high-throughput amplicon sequencing. J Virol **87:**13837-13852.
181. **Santoni FA, Guerra J, Luban J.** 2012. HERV-H RNA is abundant in human embryonic stem cells and a precise marker for pluripotency. Retrovirology **9:**111.
182. **Chhangawala S, Rudy G, Mason CE, Rosenfeld JA.** 2015. The impact of read length on quantification of differentially expressed genes and splice junction detection. Genome Biol **16:**131.
183. **Turner G, Barbulescu M, Su M, Jensen-Seaman MI, Kidd KK, Lenz J.** 2001. Insertional polymorphisms of full-length endogenous retroviruses in humans. Curr Biol **11:**1531-1535.
184. **Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D.** 2002. The human genome browser at UCSC. Genome Res **12:**996-1006.
185. **Tamura K, Stecher G, Peterson D, Filipski A, Kumar S.** 2013. MEGA6: Molecular Evolutionary Genetics Analysis version 6.0. Mol Biol Evol **30:**2725-2729.
186. **Edgar RC.** 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res **32:**1792-1797.
187. **Untergasser A, Cutcutache I, Koressaar T, Ye J, Faircloth BC, Remm M, Rozen SG.** 2012. Primer3--new capabilities and interfaces. Nucleic Acids Res **40:**e115.
188. **McWilliam H, Li W, Uludag M, Squizzato S, Park YM, Buso N, Cowley AP, Lopez R.** 2013. Analysis Tool Web Services from the EMBL-EBI. Nucleic Acids Res **41:**W597-600.
189. **Soding J.** 2005. Protein homology detection by HMM-HMM comparison. Bioinformatics **21:**951-960.
190. **Bhardwaj N, Maldarelli F, Mellors J, Coffin JM.** 2014. HIV-1 infection leads to increased transcription of human endogenous retrovirus HERV-K (HML-2) proviruses in vivo but not to increased virion production. J Virol **88:**11108-11120.
191. **Bolger AM, Lohse M, Usadel B.** 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics **30:**2114-2120.
192. **Macmanes MD.** 2014. On the optimal trimming of high-throughput mRNA sequence data. Front Genet **5:**13.
193. **Garcia-Alcalde F, Okonechnikov K, Carbonell J, Cruz LM, Gotz S, Tarazona S, Dopazo J, Meyer TF, Conesa A.** 2012. Qualimap: evaluating next-generation sequencing alignment data. Bioinformatics **28:**2678-2679.

194. **Magoc T, Salzberg SL.** 2011. FLASH: fast length adjustment of short reads to improve genome assemblies. Bioinformatics **27:**2957-2963.
195. **Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL.** 2013. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol **14:**R36.
196. **Langmead B, Salzberg SL.** 2012. Fast gapped-read alignment with Bowtie 2. Nat Methods **9:**357-359.
197. **Barbosa-Morais NL, Irimia M, Pan Q, Xiong HY, Gueroussov S, Lee LJ, Slobodeniuc V, Kutter C, Watt S, Colak R, Kim T, Misquitta-Ali CM, Wilson MD, Kim PM, Odom DT, Frey BJ, Blencowe BJ.** 2012. The evolutionary landscape of alternative splicing in vertebrate species. Science **338:**1587-1593.
198. **Ono M, Kawakami M, Takezawa T.** 1987. A novel human nonviral retroposon derived from an endogenous retrovirus. Nucleic Acids Res **15:**8725-8737.
199. **Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, Genome Project Data Processing S.** 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics **25:**2078-2079.
200. **Thorvaldsdottir H, Robinson JT, Mesirov JP.** 2013. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Brief Bioinform **14:**178-192.
201. **Cartharius K, Frech K, Grote K, Klocke B, Haltmeier M, Klingenhoff A, Frisch M, Bayerlein M, Werner T.** 2005. MatInspector and beyond: promoter analysis based on transcription factor binding sites. Bioinformatics **21:**2933-2942.
202. **Genomes Project C, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, Korbel JO, Marchini JL, McCarthy S, McVean GA, Abecasis GR.** 2015. A global reference for human genetic variation. Nature **526:**68-74.
203. **Aken BL, Achuthan P, Akanni W, Amode MR, Bernsdorff F, Bhai J, Billis K, Carvalho-Silva D, Cummins C, Clapham P, Gil L, Giron CG, Gordon L, Hourlier T, Hunt SE, Janacek SH, Juettemann T, Keenan S, Laird MR, Lavidas I, Maurel T, McLaren W, Moore B, Murphy DN, Nag R, Newman V, Nuhn M, Ong CK, Parker A, Patricio M, Riat HS, Sheppard D, Sparrow H, Taylor K, Thormann A, Vullo A, Walts B, Wilder SP, Zadissa A, Kostadima M, Martin FJ, Muffato M, Perry E, Ruffier M, Staines DM, Trevanion SJ, Cunningham F, Yates A, Zerbino DR, Flicek P.** 2017. Ensembl 2017. Nucleic Acids Res **45:**D635-D642.
204. **Manghera M, Douville RN.** 2013. Endogenous retrovirus-K promoter: a landing strip for inflammatory transcription factors? Retrovirology **10:**16.
205. **Hughes JF, Coffin JM.** 2004. Human endogenous retrovirus K solo-LTR formation and insertional polymorphisms: implications for human and viral evolution. Proc Natl Acad Sci U S A **101:**1668-1672.
206. **Kornienko AE, Guenzl PM, Barlow DP, Pauler FM.** 2013. Gene regulation by the act of long non-coding RNA transcription. BMC Biol **11:**59.
207. **Allard STM, Kopish K.** 2008. Luciferase reporter assays: Powerful, adaptable tools for cell biology research. Cell Notes **21:**23-26.
208. **Yun C, Dasgupta R.** 2014. Luciferase reporter assay in Drosophila and mammalian tissue culture cells. Curr Protoc Chem Biol **6:**7-23.

209. **Buzdin A, Ustyugova S, Khodosevich K, Mamedov I, Lebedev Y, Hunsmann G, Sverdlov E.** 2003. Human-specific subfamilies of HERV-K (HML-2) long terminal repeats: three master genes were active simultaneously during branching of hominoid lineages. Genomics **81:**149-156.
210. **Solberg N, Krauss S.** 2013. Luciferase assay to study the activity of a cloned promoter DNA fragment. Methods Mol Biol **977:**65-78.
211. **Ono M, Yasunaga T, Miyata T, Ushikubo H.** 1986. Nucleotide sequence of human endogenous retrovirus genome related to the mouse mammary tumor virus genome. J Virol **60:**589-598.
212. **Seifarth W, Baust C, Murr A, Skladny H, Krieg-Schneider F, Blusch J, Werner T, Hehlmann R, Leib-Mosch C.** 1998. Proviral structure, chromosomal location, and expression of HERV-K-T47D, a novel human endogenous retrovirus derived from T47D particles. J Virol **72:**8384-8391.
213. **Keydar I, Ohno T, Nayak R, Sweet R, Simoni F, Weiss F, Karby S, Mesa-Tejada R, Spiegelman S.** 1984. Properties of retrovirus-like particles produced by a human breast carcinoma cell line: immunological relationship with mouse mammary tumor virus proteins. Proc Natl Acad Sci U S A **81:**4188-4192.
214. **Elenbaas B, Spirio L, Koerner F, Fleming MD, Zimonjic DB, Donaher JL, Popescu NC, Hahn WC, Weinberg RA.** 2001. Human breast cancer cells generated by oncogenic transformation of primary mammary epithelial cells. Genes Dev **15:**50-65.
215. **Feng J, Funk WD, Wang SS, Weinrich SL, Avilion AA, Chiu CP, Adams RR, Chang E, Allsopp RC, Yu J, et al.** 1995. The RNA component of human telomerase. Science **269:**1236-1241.
216. **Okuda K, Bardeguez A, Gardner JP, Rodriguez P, Ganesh V, Kimura M, Skurnick J, Awad G, Aviv A.** 2002. Telomere length in the newborn. Pediatr Res **52:**377-381.
217. **Jaskelioff M, Muller FL, Paik JH, Thomas E, Jiang S, Adams AC, Sahin E, Kost-Alimova M, Protopopov A, Cadinanos J, Horner JW, Maratos-Flier E, Depinho RA.** 2011. Telomerase reactivation reverses tissue degeneration in aged telomerase-deficient mice. Nature **469:**102-106.
218. **Raynaud CM, Sabatier L, Philipot O, Olaussen KA, Soria JC.** 2008. Telomere length, telomeric proteins and genomic instability during the multistep carcinogenic process. Crit Rev Oncol Hematol **66:**99-117.
219. **Dimri G, Band H, Band V.** 2005. Mammary epithelial cell transformation: insights from cell culture and mouse models. Breast Cancer Res **7:**171-179.
220. **Rangarajan A, Hong SJ, Gifford A, Weinberg RA.** 2004. Species- and cell type-specific requirements for cellular transformation. Cancer Cell **6:**171-183.
221. **Victorino VJ, Campos FC, Herrera AC, Colado Simao AN, Cecchini AL, Panis C, Cecchini R.** 2014. Overexpression of HER-2/neu protein attenuates the oxidative systemic profile in women diagnosed with breast cancer. Tumour Biol **35:**3025-3034.
222. **Nahta R, Yu D, Hung MC, Hortobagyi GN, Esteva FJ.** 2006. Mechanisms of disease: understanding resistance to HER2-targeted therapy in human breast cancer. Nat Clin Pract Oncol **3:**269-280.

223. **Downward J.** 2003. Targeting RAS signalling pathways in cancer therapy. Nat Rev Cancer **3:**11-22.
224. **Olsson E, Winter C, George A, Chen Y, Torngren T, Bendahl PO, Borg A, Gruvberger-Saal SK, Saal LH.** 2015. Mutation Screening of 1,237 Cancer Genes across Six Model Cell Lines of Basal-Like Breast Cancer. PLoS One **10:**e0144528.
225. **Re MA, Azad RK.** 2014. Generalization of entropy based divergence measures for symbolic sequence analysis. PLoS One **9:**e93532.
226. **Mircsof D, Langouet M, Rio M, Moutton S, Siquier-Pernet K, Bole-Feysot C, Cagnard N, Nitschke P, Gaspar L, Znidaric M, Alibeu O, Fritz AK, Wolfer DP, Schroter A, Bosshard G, Rudin M, Koester C, Crestani F, Seebeck P, Boddaert N, Prescott K, Study DDD, Hines R, Moss SJ, Fritschy JM, Munnich A, Amiel J, Brown SA, Tyagarajan SK, Colleaux L.** 2015. Mutations in NONO lead to syndromic intellectual disability and inhibitory synaptic defects. Nat Neurosci **18:**1731-1736.
227. **Foti R, Gnan S, Cornacchia D, Dileep V, Bulut-Karslioglu A, Diehl S, Buness A, Klein FA, Huber W, Johnstone E, Loos R, Bertone P, Gilbert DM, Manke T, Jenuwein T, Buonomo SC.** 2016. Nuclear Architecture Organized by Rif1 Underpins the Replication-Timing Program. Mol Cell **61:**260-273.
228. **Kao SH, Wu KJ, Lee WH.** 2016. Hypoxia, Epithelial-Mesenchymal Transition, and TET-Mediated Epigenetic Changes. J Clin Med **5**.
229. **Barber RD, Harmer DW, Coleman RA, Clark BJ.** 2005. GAPDH as a housekeeping gene: analysis of GAPDH mRNA expression in a panel of 72 human tissues. Physiol Genomics **21:**389-395.
230. **Rycaj K, Plummer JB, Yin B, Li M, Garza J, Radvanyi L, Ramondetta LM, Lin K, Johanning GL, Tang DG, Wang-Johanning F.** 2015. Cytotoxicity of human endogenous retrovirus K-specific T cells toward autologous ovarian cancer cells. Clin Cancer Res **21:**471-483.
231. **Strausberg RL, Feingold EA, Grouse LH, Derge JG, Klausner RD, Collins FS, Wagner L, Shenmen CM, Schuler GD, Altschul SF, Zeeberg B, Buetow KH, Schaefer CF, Bhat NK, Hopkins RF, Jordan H, Moore T, Max SI, Wang J, Hsieh F, Diatchenko L, Marusina K, Farmer AA, Rubin GM, Hong L, Stapleton M, Soares MB, Bonaldo MF, Casavant TL, Scheetz TE, Brownstein MJ, Usdin TB, Toshiyuki S, Carninci P, Prange C, Raha SS, Loquellano NA, Peters GJ, Abramson RD, Mullahy SJ, Bosak SA, McEwan PJ, McKernan KJ, Malek JA, Gunaratne PH, Richards S, Worley KC, Hale S, Garcia AM, Gay LJ, Hulyk SW, Villalon DK, Muzny DM, Sodergren EJ, Lu X, Gibbs RA, Fahey J, Helton E, Ketteman M, Madan A, Rodrigues S, Sanchez A, Whiting M, Madan A, Young AC, Shevchenko Y, Bouffard GG, Blakesley RW, Touchman JW, Green ED, Dickson MC, Rodriguez AC, Grimwood J, Schmutz J, Myers RM, Butterfield YS, Krzywinski MI, Skalska U, Smailus DE, Schnerch A, Schein JE, Jones SJ, Marra MA; Mammalian Gene Collection Program Team.** 2002. Generation and initial analysis of more than 15,000 full-length human and mouse cDNA sequences. Proc Natl Acad Sci U S A **99:**16899-16903.

232. **Barbulescu M, Turner G, Seaman MI, Deinard AS, Kidd KK, Lenz J.** 1999. Many human endogenous retrovirus K (HERV-K) proviruses are unique to humans. Curr Biol **9:**861-868.
233. **Hughes JF, Coffin JM.** 2001. Evidence for genomic rearrangements mediated by human endogenous retroviruses during primate evolution. Nat Genet **29:**487-489.
234. **Kovalskaya E, Buzdin A, Gogvadze E, Vinogradova T, Sverdlov E.** 2006. Functional human endogenous retroviral LTR transcription start sites are located between the R and U5 regions. Virology **346:**373-378.
235. **Meyer M, Kircher M, Gansauge MT, Li H, Racimo F, Mallick S, Schraiber JG, Jay F, Prufer K, de Filippo C, Sudmant PH, Alkan C, Fu Q, Do R, Rohland N, Tandon A, Siebauer M, Green RE, Bryc K, Briggs AW, Stenzel U, Dabney J, Shendure J, Kitzman J, Hammer MF, Shunkov MV, Derevianko AP, Patterson N, Andres AM, Eichler EE, Slatkin M, Reich D, Kelso J, Paabo S.** 2012. A high-coverage genome sequence from an archaic Denisovan individual. Science **338:**222-226.
236. **Reich D, Green RE, Kircher M, Krause J, Patterson N, Durand EY, Viola B, Briggs AW, Stenzel U, Johnson PL, Maricic T, Good JM, Marques-Bonet T, Alkan C, Fu Q, Mallick S, Li H, Meyer M, Eichler EE, Stoneking M, Richards M, Talamo S, Shunkov MV, Derevianko AP, Hublin JJ, Kelso J, Slatkin M, Paabo S.** 2010. Genetic history of an archaic hominin group from Denisova Cave in Siberia. Nature **468:**1053-1060.
237. **Suntsova M, Gogvadze EV, Salozhin S, Gaifullin N, Eroshkin F, Dmitriev SE, Martynova N, Kulikov K, Malakhova G, Tukhbatova G, Bolshakov AP, Ghilarov D, Garazha A, Aliper A, Cantor CR, Solokhin Y, Roumiantsev S, Balaban P, Zhavoronkov A, Buzdin A.** 2013. Human-specific endogenous retroviral insert serves as an enhancer for the schizophrenia-linked gene PRODH. Proc Natl Acad Sci U S A **110:**19472-19477.
238. **Johanning GL, Malouf GG, Zheng X, Esteva FJ, Weinstein JN, Wang-Johanning F, Su X.** 2017. Expression of human endogenous retrovirus-K is strongly associated with the basal-like breast cancer phenotype. Sci Rep **7:**41960.
239. **Vincendeau M, Gottesdorfer I, Schreml JM, Wetie AG, Mayer J, Greenwood AD, Helfer M, Kramer S, Seifarth W, Hadian K, Brack-Werner R, Leib-Mosch C.** 2015. Modulation of human endogenous retrovirus (HERV) transcription during persistent and de novo HIV-1 infection. Retrovirology **12:**27.
240. **Etkind PR, Lumb K, Du J, Racevskis J.** 1997. Type 1 HERV-K genome is spliced into subgenomic transcripts in the human breast tumor cell line T47D. Virology **234:**304-308.
241. **Schmitt K, Heyne K, Roemer K, Meese E, Mayer J.** 2015. HERV-K(HML-2) rec and np9 transcripts not restricted to disease but present in many normal human tissues. Mob DNA **6:**4.
242. **Domansky AN, Kopantzev EP, Snezhkov EV, Lebedev YB, Leib-Mosch C, Sverdlov ED.** 2000. Solitary HERV-K LTRs possess bi-directional promoter activity and contain a negative regulatory element in the U5 region. FEBS Lett **472:**191-195.

243. **Cullen BR, Lomedico PT, Ju G.** 1984. Transcriptional interference in avian retroviruses--implications for the promoter insertion model of leukaemogenesis. Nature **307:**241-245.
244. **Gibb EA, Warren RL, Wilson GW, Brown SD, Robertson GA, Morin GB, Holt RA.** 2015. Activation of an endogenous retrovirus-associated long non-coding RNA in human adenocarcinoma. Genome Med **7:**22.
245. **Shah N, Sukumar S.** 2010. The Hox genes and their roles in oncogenesis. Nat Rev Cancer **10:**361-371.
246. **Tammimies K, Bieder A, Lauter G, Sugiaman-Trapman D, Torchet R, Hokkanen ME, Burghoorn J, Castren E, Kere J, Tapia-Paez I, Swoboda P.** 2016. Ciliary dyslexia candidate genes DYX1C1 and DCDC2 are regulated by Regulatory Factor (RF) X transcription factors through X-box promoter motifs. FASEB J doi:10.1096/fj.201500124RR.
247. **Jiang S, Zhang E, Zhang R, Li X.** 2016. Altered activity patterns of transcription factors induced by endoplasmic reticulum stress. BMC Biochem **17:**8.
248. **Cook DN, Kang HS, Jetten AM.** 2015. Retinoic Acid-Related Orphan Receptors (RORs): Regulatory Functions in Immunity, Development, Circadian Rhythm, and Metabolism. Nucl Receptor Res **2**.
249. **Zaret KS, Carroll JS.** 2011. Pioneer transcription factors: establishing competence for gene expression. Genes Dev **25:**2227-2241.
250. **Magnani L, Eeckhoute J, Lupien M.** 2011. Pioneer factors: directing transcriptional regulators within the chromatin environment. Trends Genet **27:**465-474.
251. **Magnani L, Ballantyne EB, Zhang X, Lupien M.** 2011. PBX1 genomic pioneer function drives ERalpha signaling underlying progression in breast cancer. PLoS Genet **7:**e1002368.
252. **Xiao Y, Ye Y, Zou X, Jones S, Yearsley K, Shetuni B, Tellez J, Barsky SH.** 2011. The lymphovascular embolus of inflammatory breast cancer exhibits a Notch 3 addiction. Oncogene **30:**287-300.
253. **Kikugawa T, Kinugasa Y, Shiraishi K, Nanba D, Nakashiro K, Tanji N, Yokoyama M, Higashiyama S.** 2006. PLZF regulates Pbx1 transcription and Pbx1-HoxC8 complex leads to androgen-independent prostate cancer proliferation. Prostate **66:**1092-1099.
254. **Park JT, Shih Ie M, Wang TL.** 2008. Identification of Pbx1, a potential oncogene, as a Notch3 target gene in ovarian cancer. Cancer Res **68:**8852-8860.
255. **Plowright L, Harrington KJ, Pandha HS, Morgan R.** 2009. HOX transcription factors are potential therapeutic targets in non-small-cell lung cancer (targeting HOX genes in lung cancer). Br J Cancer **100:**470-475.
256. **Morgan R, Pirard PM, Shears L, Sohal J, Pettengell R, Pandha HS.** 2007. Antagonism of HOX/PBX dimer formation blocks the in vivo proliferation of melanoma. Cancer Res **67:**5806-5813.
257. **Jozwik KM, Carroll JS.** 2012. Pioneer factors in hormone-dependent cancers. Nat Rev Cancer **12:**381-385.

258. **Morgan R, Boxall A, Harrington KJ, Simpson GR, Gillett C, Michael A, Pandha HS.** 2012. Targeting the HOX/PBX dimer in breast cancer. Breast Cancer Res Treat **136:**389-398.
259. **Quail MA, Smith M, Coupland P, Otto TD, Harris SR, Connor TR, Bertoni A, Swerdlow HP, Gu Y.** 2012. A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. BMC Genomics **13:**341.
260. **Iwafuchi-Doi M, Zaret KS.** 2014. Pioneer transcription factors in cell reprogramming. Genes Dev **28:**2679-2692.