ENDOGENOUS RETROVIRUSES DRIVE TRANSCRIPTIONAL INNOVATION IN HUMAN CANCER

by

Artem Babaian

B. Sc., McMaster University, 2012

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF

THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

in

THE FACULTY OF GRADUATE AND POSTDOCTORAL STUDIES

(MEDICAL GENETICS)

THE UNIVERSITY OF BRITISH COLUMBIA

(Vancouver)

August 2019

The following individuals certify that they have read, and recommend to the Faculty of Graduate and Postdoctoral Studies for acceptance, the dissertation entitled:

Endogenous Retroviruses Drive Transcriptional Innovation In Human Cancer

submitted by Artem Babaian in partial fulfillment of the requirements for
the degree of Doctor of Philosophy
in Medical Genetics

Examining Committee:

Dixie Mager
Supervisor

Carolyn Brown
Supervisory Committee Member

Matt Lorincz
Supervisory Committee Member

Inanc Birol
University Examiner

Paul Pavlidis
University Examiner

Ting Wang
External Examiner

Additional Supervisory Committee Members:

Martin Hirst
Supervisory Committee Member

Wyeth Wasserman
Supervisory Committee Member

Abstract

Transposable element (TE) exaptation is the process of TE incorporation into functional, and in some cases necessary, genes or regulatory units over evolutionary time. I postulate that an analogous process occurs in oncogenesis, wherein TE-derived promoters generate “noisy” transcription and novel transcripts which can then undergo selection to drive cancer transcriptome evolution. Such “*onco-exaptation*” is reviewed in the context of several cancers including Hodgkin Lymphoma (HL) where it results in expression of the oncogene *CSF1R*, yet it is unclear how widespread this phenomenon is.

I hypothesize that epigenomic dysregulation in cancer leads to a genome-wide derepression of TE-initiated transcripts, some of which have an oncogenic role. To address this hypothesis, I developed a computational tool called ‘*LIONS*’ to analyze RNA-sequencing data for TE-initiated transcripts. *LIONS* detects and quantifies TE-initiated transcripts through transcriptome assembly, applies a novel artificial neural network classifier to identify TE promoter events, and compares biological sets of data.

Using this tool, I have determined that the transcriptomes of colorectal carcinoma, diffuse large B-cell lymphoma and HL all have an overall increase in TE-initiated transcripts relative to their respective controls. This increase is specifically driven by an increase in endogenous retroviral long terminal repeat (LTR) initiated transcripts. The distribution of this TE transcriptional activity is widely distributed across the genome, yet patterns of co-activation among element families and the recurrent activation of a small sub-set of TEs is evident.

One such recurrent TE-initiated transcript is the LOR1a LTR driven expression of the *IRF5* oncogene in HL. *IRF5*, along with *CSF1R* and a panel of putative oncogenic TE-initiated transcripts were explored as novel biomarkers in HL. Altogether, I propose that the process of onco-exaptation

is a novel and distinct mechanism for oncogene activation and a model system for future studies of exaptation and transcriptome evolution.

Lay Summary

Half of all human DNA is made up of self-replicating ‘jumping genes’ called transposable elements (TEs), and near 20% of TEs come from ancient retroviral infections (viruses of the same type as Human Immunodeficiency Virus). In this thesis I describe how these ancient viral gene components, which are normally turned off by our cells, can become re-activated in human cancers, and re-used for the inappropriate expression of cancer-promoting genes. This research helps further understand how cancer cells develop an inappropriate gene expression profile and may help in the development of novel bio-medical applications.

Preface

The dissertation is the original intellectual product of the author, A. Babaian, unless otherwise noted below.

Use of primary human RNA-seq data was covered by human ethics certificate H14-01561 and laboratory work was conducted under biosafety certificate B17-0213, both approved by the University of British Columbia.

A version of Chapter 1 section 1.2 and Chapter 5 has been published as a review article written collaboratively by D. L. Mager and myself as Babaian, A. & Mager, D. L. Endogenous retroviral promoter exaptation in human cancer. *Mob DNA* **7**, 24 (2016).

A version of Chapter 2 has been published as Babaian, A., Thompson, R., Lever J., Gagnier, L., Karimi, M.M., Mager, D.L. LIONS: Analysis suite for detecting and quantifying transposable element initiated transcription from RNA-seq. *Bioinforamtics* **btz130**, (*2019).* I was the lead developer for *LIONS* software development, J. Lever wrote the Python read annotation program and R. Thompson wrote the Docker installation files. The RT-PCR data for Figure 2.8 was generated by L. Gagnier and used with permission.

A version of Chapter 4 has been published as Babaian, A. *et al.* Onco-exaptation of an endogenous retroviral LTR drives IRF5 expression in Hodgkin lymphoma. *Oncogene* **35**, 2542–2546 (2016). M.T. Romanish originally identified *IRF5* as an LTR-derived chimeric transcript based on work by M.M. Karimi. L. Gagnier performed the 5’ RACE and bisulphite sequencing experiments in Figure 4.3. L.Y. Kuo performed Western blotting in Figure 4.4. The Hodgkin Lymphoma microarray data was kindly provided by C. Steidl from Steidl, C. *et al.* Gene expression profiling of microdissected Hodgkin Reed-Sternberg cells correlates with treatment outcome in classical Hodgkin lymphoma. *Blood* **120**, 3530–3540 (2012).

Table of Contents

List of Tables

List of Figures

List of Supplementary Materials

Acknowledgments

I am indebted to the teachers and mentors who have helped shaped me as a scientist; my brother George, Mr. Robert Smachylo, Ms. Emily Grant, Dr. Roger Jacobs, Dr. Marie Elliot, Dr. Markus Czub, Mr. Sampson Law, Dr. Ali Ashkar, and Dr. Gregg Morin; as well as the communities at the Terry Fox Laboratory and Department of Medical Genetics.

I'd like to thank my committee members, Carolyn Brown, Martin Hirst, Matt Lorincz, and Wyeth Wasserman for their patience and continued guidance.

I can't possibly overstate the influence of my partner, Katharina Rothe, for her love and companionship over years to bring me to this point.

Finally, I'd like to thank my supervisor Dixie Mager for letting me join her lab, sharing her knowledge and making this work possible.

Dedication

for Alfia

Chapter 1: A primer on human transposable elements

Transposable elements (TEs) make up at least half of the human genome [1,2]. If one were to stretch the definition of an organism, TEs can be imagined as organisms living within the ecosystem of the host DNA, and encoding the information necessary to undergo the function of transposition (mobilization) within that environment [3–5].

There are two categories of TEs; DNA transposons which directly transpose using a transposase enzyme (cut and paste mechanism); and the retrotransposons which transpose via a transcribed RNA intermediate and the reverse transcriptase enzyme (copy and paste mechanism). The retrotransposons are long terminal repeat (LTR) elements, long interspersed repeat elements (LINEs) and short interspersed repeat elements (SINEs) including complex elements such as the SINE-R VNTR Alu (SVA) composite element [6].

The vast majority of human TEs are non-active remnants of their ancient ancestors. TEs are mutated, fragmented and actively repressed such that their capacity to complete the transposition program is lost, yet sub-functions in this program, such as gene regulation or the production of a protein product, may remain intact. In rare instances through evolution, these TE sub-functions meld into a host program through a process called *exaptation* [7,8].

TEs have been referred to as “Junk DNA,” which is read to mean that TEs are *garbage* with no consequence to an organism’s biology except as a detrimental parasite [9]. An alternative framework views the “Junk DNA” term more neutrally, namely that TEs are a reservoir of potentially useful genetic sequences which have no immediate function, like an old bicycle in a junk-yard. The presence of TEs though allows for the possibility for evolutionary changes to ‘tinker’ [10,11] in the creation of novelty. Thus, in an evolving system, TEs can be thought to

increase the rate of adaptation by increasing genetic diversity with functionally competent DNA sequences.

Cancer is fundamentally an evolutionary disease. A cell functioning in the context of a co-operative multi-cellular organism is transformed through evolutionary/selective forces into a cell undergoing selfish proliferation. The genome and epigenome of cancer cells adapt in this way and novel molecular functions arise. My central premise is that TEs are a reservoir of genetic innovation in human cancer, and their transcriptional capacity is co-opted to accelerate tumorigenesis.

1.1 Human transposable elements

In the human reference genome, TEs make up at least 47.8% of the total non-N (a known A, T, C or G base) sequence, although this is likely an underestimate as sequence divergence shrouds their identification. LINEs are the most abundant class of TE making up 21.8% of the genome, followed by the SINEs (13.4%), LTRs (9.12%), and DNA (3.45%) (Table 1.1 and [1,2,12]). The origins, molecular mechanisms for replication and consequences on the host genome biology varies for each of these elements.

	Genome Size	Non-N Genome
h38	3,209,286,105	3,049,315,783

	Basepairs	% Genome	% TE
RepeatMasker Total	1,456,392,972	47.76	100.00
LINE	663,376,262	21.75	45.55
SINE	409,836,828	13.44	28.14
LTR	277,960,976	9.12	19.09
DNA	105,218,906	3.45	7.22

Table 1.1: Genomic abundance of human transposable elements
Summary statistics for the abundance of each TE family as annotated in RepeatMasker [2] for the hg38 human reference genome.

1.1.1 Endogenous retroviruses and long terminal repeats

Endogenous retroviruses (ERVs), as their name implies, are retroviruses that have integrated into the host germ-line. Infection and integration into the DNA of germ-line cells means that the viral DNA is transmitted from parent to offspring (vertical transmission). This is in contrast to the horizontal transmission of *exogenous* retroviruses (XRVs) where the virus infects the somatic cells of the host and must be transmitted to a new host to propagate.

Phylogenetically, human ERVs (HERVs) are recognized as being derived from retroviruses, in particular the Orthoretrovirinae subfamily, since they share a homologous genetic organization (Figure 1.1 and Table 1.2) with the exception of the rare and ancient *Gypsy* elements, which are the only non-ERV LTR retrotransposons in mammals [13]. The prototypical proviral genome contains three genes; *gag* a structural capsid component, *pol* the reverse transcriptase, and *env* the envelop protein. The *gag-pol-env* genome is flanked by identical gene regulatory sequences called Long Terminal Repeats (LTRs).

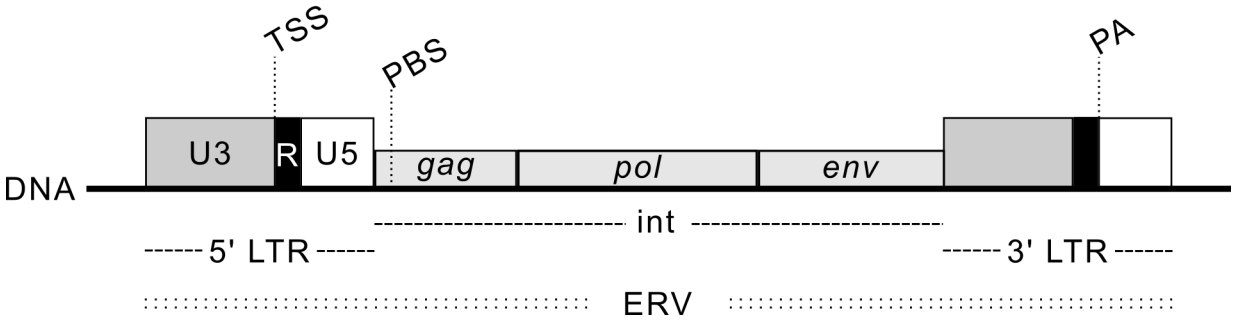


Figure 1.1: Genetic organization of a prototypical retrovirus integrated in a host genome

A prototypical schematic of provirus in the host genome. Two flanking long terminal repeats (LTRs) flank the internal sequence of the retrovirus (int). Each LTR is sub-divided into the U3, R and U5 regions, defined by the transcription start site (TSS) and poly-adenylation site (PA). Endogenous retroviruses (ERVs) are often classified by the tRNA complementary of their primer binding sites (PBS). [351]

ERV Class	LTR (%)	ERV Group	LTR Names*	Max Coverage of Internal Regions	Max Coverage of LTRs	Est. % Solitary LTRs**
I: ERV1	31.7	HERV-9	LTR12, 12B - 12F	454	7439	93.50
		HERV-E	LTR2, 2B, 2C	246	1206	74.38
		HERV-H	LTR7, 7B, 7C, 7Y	1058	3384	54.51
		HUERS-P2	LTR1, 1B - 1D	121	3214	96.09
		LOR1	LOR1a, 1b	175	1426	86.01
		MER41	MER41A - 41G	275	4659	93.73
II: ERVK	3.48	HERV-K	LTR5, 5B, 5_Hs	82	1300	93.27
		HERV-KC4, HERV-K14	LTR14, 14A - 14C	233	620	39.79
III: ERVL	21.4	HERV-L	MLT2A - 2F	793	20456	95.97
		HERV-16	LTR16A - 16E	856	19808	95.48
III: ERVL-MaLR	41.2	THE1	THE1A - 1D	7893	37043	72.92
		MLT1	MLT1A - 1O	3823	154196	97.46
IV: Gypsy	1.8					

Table 1.2: Copy number estimate of representative LTR retrotransposons

The copy number of some groups of human endogenous retroviruses was estimated from the maximum non-redundant coverage reached in the alignment of Dfam hidden Markov model alignments [11] in the hg38 reference genome. ***** *Repeat names follow RepBase [12] nomenclature.* ****** *The percent solitary LTRs were estimated assuming two LTRs are associated with each internal LTR sequence. The formula is:*

$$100 * [(number\_LTR - 2 * number\_Internal) / (number\_LTR - number\_Internal)]$$

The relationship between XRV and ERV can be fluid, especially among very young and recently integrated ERVs. Notably in *Phascolarctos cinereus* (the koala) of Australia, Koala Retrovirus (KoRV) which entered the species in the last 200 years, is an XRV in some koala populations, an ERV in others, and absent still from some island populations [14]. As well, in mice, which have high ERV activity, Murine Leukemia Virus (MuLV) undergoes both endogenization and exogenization [15].

Upon endogenization, some of the components necessary for the production of an exogenous infectious viral particle become superfluous to ERV replication, and presumably as an evolutionary optimization, the internal ERV (ERV-int) sequence can reduce to LTR-gag-pol-LTR. As such the

ERV can also be called an LTR retrotransposon. In time, the ERV sequences mutate in their host genomes and recombination between the homologous 5' and 3' LTRs leads the excision of the ERV-int sequence, resulting in a single, solitary LTR. In humans and other mammals, these solitary LTRs constitute the major fraction of ERV-derived sequence [16,17].

The degradation of the components to produce infectious viral particles does not necessarily preclude ERV replication, instead ERVs specialize for intra-cellular replication, copying and pasting their reduced genomes with higher efficiency and increasing the number of copies per genome [18,19]. Thus over evolutionary periods of time, a lineage infected with an active ERV accumulate these pathogens through iterative copy-paste, copy-paste, and copy-paste retrotranspositions. ERV amplification is balanced by mutational forces which degrade the ERV sequences leading to loss of retrotransposition activity, natural selection against organisms with a high detrimental load of ERVs, and host adaptation to actively suppress retrotransposition or shifts in the molecular biology of an organism to no longer support retrotransposition (see section 1.1.8 Host-pathogen co-evolution and exaptation).

The genomic ERV load and ERV (retrotranspositional) activity varies substantially across species. For instance, mice have high ERV retrotransposition, with 10-12% of spontaneous phenotype-causing mutations arising from an ERV insertion [20,21], while the human pan-genome is notably depleted of ERV transpositions [22]. Indeed, no modern insertions of ERVs have been documented in humans [22–24].

The 9.12% of the human genome derived from ERVs range from 50+ million years before present (Ma) to the most recent human ERV retrotranspositions (of the HERV-K HML-2 elements) occurring as late as 100,000 years before present [25]. ERVs are waning in the human genome, from the ~540,000 ERV sites in the human genome, 81.9% are now solitary LTRs with no new insertions replenishing them. The ancestral function of these LTRs is to enhance and initiate

transcription for the viral genome. As such LTRs are a reservoir of dormant regulatory sequences dispersed across the genome.

1.1.2 Long interspersed repeat elements (LINEs)

LINEs are the most abundant class of TEs in the human genome by number of bases. However, while there are ~1,500,000 LINE fragments in the genome, as little as 80-100 of them (of the more recent LINE-1 (L1) family) remain active and potentially able to autonomously retrotranspose [26,27]. L1 initiates from an internal Polymerase II promoter (~900 bp on the 5' end) and contains three ORFs; *ORF1* which binds LINE mRNA, *ORF2* which encodes for the endonuclease and reverse transcriptase and the recently described LINE-1 *ORF0* encoded in the antisense direction which promotes L1 transposition [28].

The common human L1 propagates via a target primed reverse transcriptase mechanism (TPRT). Upon transcription, LINE mRNA is translated into ORF1p and ORF2p which associate back to the mRNA *in cis.* This ribonucleoprotein complex is then transported to the nucleus where the ORF2p dimer nicks the host DNA near a TTAAAA motif which is used to prime the ORF2p-mediated reverse transcription. This process is often prematurely aborted leading to 5` truncation of the new LINE (and consequently loss of the promoter sequence) [29]. There are ~0.03–0.125 germline L1 transposition events per human generation [26,30].

One exciting and controversial area of research in LINE biology is the finding that L1 generates somatic genome mosaicism in the human brain [31–33]. It is certain that bona fide L1 insertions can be detected by amplification-based methods from brain necropsy samples, what is debated is the exact frequency of this occurrence. The rates are estimated to be between 0.04 and 13.7 L1 insertions per adult neuron and at ~90 billion neurons there is a conservative 3.6 billion somatic brain L1 insertions accumulated over the lifetime of an individual [34,35]. In addition to being a source of mutations, this research raises fundamental questions regarding the function, if any, that

somatic mosaicism plays in human physiology. Is it possible to consider a tissue lineage as undergoing genetic or epigenetic specialization? A question which will be re-visited throughout this thesis.

1.1.3 Short interspersed repeat elements (SINEs)

SINEs, specifically the Alu elements, are one of the most successful and active human mobile elements having >1 million copies (or ~1.8 million fragments) per haploid genome. SINEs are typically between 100 – 600 bp and have arisen multiple independent times in evolution, originating from pseudogenes of other RNA species such as: transfer-RNA (tRNA), signal recognition RNA (7SL), 5S ribosomal RNA (5S rRNA), 28S rRNA, or small nucleolar RNA (snRNA), and also often incorporate DNA from LINEs or other sequences of unknown origin [36–38]. What distinguishes SINEs from other TEs is that it is transcribed by Polymerase III and requires exogenous reverse transcriptase (typically ORF2p from LINEs) for its transposition [36].

The canonical human Alu element, derived from a truncated 7SL RNA [39] is ~280 nt in length. Transcription initiates from an internal PolIII promoter (and is terminated by a non-Alu, adjacent genomic poly-T). The folded Alu RNA then associates with two heterodimers of signal recognition protein (SRP) 9 and 14; and Poly-A Binding Protein [40,40]. This Alu RNP then recruits LINE1 ORF2p for its retrotransposition [41]. There is a novel germline SINE transposition event every ~20 human births [42].

Besides the Alu elements, there are also older and non-transpositional Mammalian-wide Interspersed Repeat (MIRs) and the recently evolved and active composite SVAs. Just over a fifth of SINEs are MIRs which are an ancient tRNA-derived elements active at least 130 Ma [43]. To this day, MIRs remain enriched for transcription factor binding sites and enhancer function as defined by chromatin markers suggesting they have been conserved in the regulatory evolution of mammals [44]. In contrast, the <15 Ma SVA elements are not known to have integrated into ‘normal’ genetic

programs. Instead the 3733 fragments (0.13% of the total) in the haploid genome [12], are retrotranspositionally active [45] and have been documented to cause human disease [46].

The mutagenic potential of SINE and LINE retrotransposition is evident from the number of events resulting in human disease, 76 Alu, 29 LINE, and 12 SVA independent events (reviewed in [46]). Noticeably, there are no known novel ERV transpositions or polymorphic insertions resulting in human disease, but retrotransposition is not the only mechanism by which TEs influence the genome.

1.1.4 DNA elements

LINEs, SINEs, and LTR retrotransposons are all Class I transposons in eukaryotes, those that transpose though an RNA intermediate using a reverse transcriptase. Class II or DNA transposons do not use an RNA intermediate, and thus lack reverse transcriptase [13]. Making up 3.4% of the human genome, DNA transposons were active as late as the primate and eutherian common ancestor (64 – 150 Ma), with tapering activity by the divergence of prosimians and new world monkeys (~40 Ma) [47]. Unfortunately, very little is known about the role of DNA elements in primate lineages, it is noteworthy that DNA elements were instrumental in the evolution of adaptive immunity 500 Ma (section 1.1.8).

In contrast to the waning of DNA transposons in most mammals a recent invasion of hATs in bats has occurred in the last ~5 Ma, coinciding with (and possibly driving) a radiation of bat species [48,49].

1.1.5 The genomic impact of TEs

Besides being neutral or detrimental to an organism, mutation and variation are an absolutely necessary (and thus beneficial) component of evolution. In a similar vein, a novel TE transposition is a mutational event that can be detrimental, neutral or beneficial to the host organism. There is

clear evidence for purifying selection against TE insertions that disrupt gene transcription [50,51], gene regulation [52,53] or even those that confer a high energetic burden associated with highly active elements [54]. Nonetheless, the high abundance of TEs in the genomes of many mammals [50,55] suggests that insertions are often near neutral, or quickly become neutral to be maintained and possibly fixed by genetic drift. In rare instances, TE insertions are beneficial to the host either as novel regulatory elements of host genes or as the origin of an entirely new gene (reviewed in [56,57]). Therefore, much like other sources of mutation, the cost of TE insertions may be off-set by the rich regulatory sequences that TEs disperse, leading to a greater rate of adaptation or even to drive speciation [58,59].

1.1.6 TE regulatory activity

Transposable elements require host transcriptional and translational machinery for their transpositional activity. This means that they must contain regulatory sequences compatible with the host genome that may include enhancers, repressors, insulators, promoters, splice acceptors and donors, and termination sequences [56]. In addition, the host genomes have evolved to suppress TE activation and thus TEs must evolve to bypass this suppression resulting in a host-pathogen arms race [60,61].

Endogenous retroviruses, LTR retrotransposons and solitary LTRs share an important regulatory structure, the LTR (Figure 1.1). Integrated retroviral elements are flanked by two identical LTRs (5’ and 3’ LTR) which after insertion, regulate the subsequent transcription of ERV RNA or mRNA. Each LTR can be sub-divided into three regions. The U3 region canonically contains the enhancer and promoter sequences necessary for ensuring the local chromatin environment is open and favorable for transcription and to initiate PolII mediated transcription of the ERV genome. The R region defines the boundaries of the ERV RNA genome, on the 5’ LTR it demarcates the transcription start site and on the 3’ LTR its terminus demarcates the transcriptional termination site.

The U5 region is not canonically required to contain regulatory sequences, but viral optimization often leads to enhancers or other sequences which increase RV/ERV efficiency to accrue in the U5 [62–64]. Internal to the 5' LTR a typical retrovirus will also contain a primer binding site (PBS) which is complementary to a host tRNA and required to prime the reverse transcription of the RNA genome. In the comprehensively studied HIV-1, over 40 mRNA isoforms are present in an infected cell which requires 4 splice acceptors and 8 splice donors [65], exemplifying the intricate regulatory structures which can exist in the ERV-int sequences.

The broad dispersion of TE regulatory sequences is believed to potentiate gene regulatory innovation [66]. In human pluripotent stem cells, a tissue tropic for many TEs, 20.9% and 14.6% of the bindings sites of the master transcription factors OCT4 and NANOG, respectively, are within a TE boundary (with a strong enrichment for ERVs) [67]. In contrast, the primate-specific Alu elements harbor ~7.5% of p53-binding sites, that require CG → TG transition from the consensus Alu to create the binding-site motif [68]. The highly abundant Alu elements also can drive species-specific alternative splicing since they contain an anti-sense cryptic splice site which may interfere with normal exon selection [69]. These examples illustrate how TEs intrinsically contain regulatory and proto-regulatory sequences which are seeded across the genome, and can act as a substrate for regulatory innovation.

1.1.7 TE transcriptional activity

Transcriptional activity of TEs is the initiation of autonomous transcription within the boundary of the element. TE transcriptional activity can be broadly divided into *ancestral transcription*, the transcription that was necessary for the mobilization of the TE as part of its original life-cycle; and *novel transcription*, the transcription initiating in the TE but at positions that are not conserved and/or consistent with the biological program of the TE. For example, transcription initiating in the R region of a 10 Ma ERV LTR is consistent with its ancestral transcription, while transcription

initiating in the center of the *env* gene (and where homologous ERVs do not share such a transcription start site (TSS)) is likely to be novel transcription. In contrast to ERVs, Alu elements cannot autonomously initiate Pol II transcription, since the Alu and the 7SL RNA from which it is derived, are both Pol III transcripts. Alu happen to be extremely abundant in humans, and this abundance leads to a high variability and ultimately novel transcription. Both ancestral and novel transcription can influence the host transcriptome; ancestral since the associated motifs and TF binding sites have evolved for initiation, and novel since the sequences involved are (in general) abundant and a substrate for spontaneous initiation to occur/evolve within.

It is important to distinguish TE autonomous transcriptional activity from TE expression, most simply measured as steady-state TE transcript levels. TE expression does not require that the TE is transcriptionally active (initiates its own transcription). A TE can be expressed in a transcript via an unrelated upstream promoter. The incorporation of non-coding sequence (such as TEs) into a mature transcript is termed 'exonization' [70–72]. Thus when measuring TE expression, limited conclusions can be drawn regarding the causative role of the TE itself in influencing the transcriptome, as TE expression alone is a correlative effect. For example, a recent study has shown that 99% of transcripts containing L1 sequences are not due to autonomous L1 transcription [73]. As well, TE fragments, particularly Alu sequences, are commonly found in the 3' UTR of coding genes [74].

Repeat sequences account for between 6 - 30% of all TSSs in the mouse and human transcriptome, depending on the tissue [75]. In humans, embryonic tissues show the most pronounced TE transcriptional activity, with ~18% of TSSs initiating in a TE when measured by cap analysis of gene expression (CAGE) [75]. Relative to non-TE initiating transcripts, the set of TE-initiated transcripts are on average expressed at a lower level [75], are enriched for long non-coding (lnc) transcripts (or depleted for protein-coding transcripts) [76,77], and show higher tissue

specificity [78]. Overall the data support a picture of TE-initiated transcripts as pervasive and hyper-variable at the levels of inter-cellular, -tissue, -individual and -species variation.

1.1.8 Host-pathogen co-evolution and exaptation

Analogous to how the immune system protects an organism from a pathogen, host genomes have evolved control factors to stave off TE expression and suppress TE transposition. Briefly, host mechanisms of TE repression include transcriptional repression via DNA methylation [79–81], or repressive histone deposition which can be targeted by Krüppel-associated box domain-zinc finger proteins (KRAB-ZFPs) [61,82–85]. TE transcripts can be targeted for degradation via APOBEC proteins [86] and through the piwi-interacting RNA (piRNA) pathway [60,87–89].

In response, TEs evolve to overcome host control factors leading to further host controls in a continual arms-race [90]. The irony is that loss of the highly evolved host control factors is lethal to both host and TE, as loss of repression results in rampant TE transcription, which is energetically costly and can lead to an intolerable load of DNA damage [87].

The vast majority of fitness altering mutations are deleterious, yet mutations are necessary for adaptation through rare advantageous mutations. TE insertions are no different, they are largely detrimental but can incorporate into the functional circuitry of an organism. The re-use of genetic elements into a novel function that confers a fitness advantage is called *exaptation* [7], and TE exaptation has shaped the natural history of many species, including humans.

One of the most substantial increases in complexity of the immune system was due to the genesis of adaptive immunity in jawed vertebrates 500 Ma [91]. Unlike older intrinsic and innate immunity, adaptive immunity generates a vast genetic diversity of antigen receptors to recognize novel non-self molecular patterns, and ‘remember’ this pattern upon subsequent encounters. This genetic ‘memory’ is created by V(D)J-recombination which allow for the generation of antigen specific B-cell receptors (and antibodies) and T-cell receptors. As early as 1979, the inverted repeat

sequence between joined V(D)J segments suggested a DNA transposon mediated mechanism [92]. This proved to be prophetic as the enzymes which mediate V(D)J recombination, RAG1 and RAG2, were shown to be homologous to the transposase of a DNA TE [93,94]. Intriguingly, the ProtoRAG DNA transposon is found in the lamprey, a jawless vertebrate, which implies there is an approximate 50 million year window in which the TE infected the genome and was exapted into the early adaptive immune system [94].

Another fascinating case-study of exaptation is the co-option of ERV *env* genes in placental mammals (including marsupials). The retroviral envelope gene has intrinsic fusogenic activity to fuse the viral membrane with the host cell membrane and allow capsid release into the cytoplasm. Young ERVs obviously contain an *env* gene, a gene often lost as the retrotransposition ERV life-cycle typically does not require extra-cellular release and infection. The placenta arose ~ 200 Ma from egg-laying mammals and ERV *env* (called *syncytins* in mammals) were co-opted and expressed in this novel structure, ultimately becoming a necessary gene, at least in mice [95]. It is hypothesized that the fusogenic and immuno-suppressive activity of the *syncytins* (mammalian exapted *env*) supports internal fetal development and inhibits the immune destruction of the developing embryo [96,97]. What is perhaps the most striking is that in the 200 Ma history of the placenta, there were at least 10 independent *syncytin* exaptation events across 7 clades, strong evidence for convergent evolution [97]. Remarkably, it was recently shown that a viviparous placental lizard also carries an ERV-derived gene with possible syncytin-like function [98], providing further support for the theory that evolution of the placental structure benefits from exaptation of viral env genes.

Exaptation of *RAG1,2* and the numerous *Syncytin* genes represents complex and significant yet rare events that shaped the natural history of our lineage. The re-use of gene regulatory sequences within TEs into already existing or emerging genes is orders of magnitude more common

[66,99,100]. There are a myriad of ways a TE can and does alter gene structure or regulation (reviewed in [66]), including in the pathogenesis of diseases such as cancer.

An example of regulatory exaptation is human *IL2RB* gene, it is expressed in the placenta through a novel THE1D LTR promoter, and not the native hematopoietic promoter [101]. Human corticotropin releasing hormone (CRH) is a biomarker for birth timing, and placenta-specific expression is specific to anthropoid primates. This species-specific expression is controlled by a THE1B LTR element, functioning not as a promoter but as an enhancer [102]. These two cases highlight a broader trend, LTRs function as species-specific promoters and enhancers in the placenta [99,103].

One important distinction between TEs and other sources of gene regulatory variability, is that repetitive element sequences are widely distributed across the genome. This means that TEs are not restricted to rewiring one locus at a time, but can have dispersed effects, altering many loci at once by providing a common response element [68]. Systems based genetic research is still in its infancy, but already there is strong evidence for broad transcriptomic alterations resulting from TEs, specifically in the dispersion of functional interferon-gamma response elements by the MER41B LTR [104,105].

The significance of the regulatory impact of TEs cannot be overstated. Indeed, it was the gene regulatory capacity of the *As-Ds* elements on the *C* gene which first led Barbera McClintock to designate them as “controlling elements”, and only later it was discovered they are the “transposable elements”, as we know them today [106,107].

1.2 Transposable elements in cancer

At it’s etiological core, cancer is an evolutionary disease. Cells which are a component part of a larger organism gain the capacity for dysregulated growth, independent of their function in the

organism. Analogous to organismal evolution, these cells are not an insular mass, they exist in a complex and dynamic environmental landscape with which they interact and must respond to, commonly referred to as the tumor microenvironment [108]. As such, the mutagenic function of TEs, in particular somatic insertions of LINEs and SINEs, has been widely explored in human oncogenesis. More recently, the idea that TEs shape changes in the regulatory landscape of cancer has emerged and is the primary topic of this thesis.

1.2.1 LINE and SINE mutagenesis in cancer

While nearly all L1s are defective, a few hundred retain the ability to retrotranspose [26] and can occasionally cause germ line mutations [24,46,109]. Several studies have documented somatic, cancer-specific L1 insertions [110–117], and a few such insertions were shown to contribute directly to malignancy [46]. For example, two L1 insertions were documented to disrupt the tumor suppressor gene *APC* in colon cancer [110,117], or the *PTEN* gene in endometrial cancer [118].

Given a gene-proximal L1 insertion, the potential for a mechanistic or regulatory impact on the gene by the L1 insertion is high. The ~6 kb or less L1 insertion contains promoters, splice acceptors and donors, and poly-A termination signals which makes the insertion more likely to knockout or "break" [119] the gene relative to physical or chemical mutagens which predominantly alter single bases. However, the rate of human L1 retrotransposition is low, with an estimated ~1-10 somatic insertion per cell lineage per human life [113]. This is in contrast to the point mutation rate of 1.45 x 10^{-8} per base per generation (~47 mutations per generation) [120] . Thus, it is probable that the overall effect size of L1 insertions on phenotype is limited as recently discussed by Hancks and Kazazian [46] along with the biological effects of LINE retrotransposition on oncogenesis. Interestingly, while Alu insertions have caused more human disease than LINEs [46], they are underrepresented relative to LINE insertions in cancers [121].

1.2.2 Retroviruses and ERVs in oncogenesis

There is currently no reported evidence for retrotranspositionally active ERVs in humans [22–24], so it is improbable that ERVs activate oncogenes or inactivate tumor suppressor genes by somatic retrotransposition. This is in contrast to the frequent oncogene activation by insertions of exogenous and endogenous retroviruses in other species or in experimental systems [122].

In seminal research on oncogenesis, Peyton Rous determined that a chicken sarcoma could be serially transplanted. Further, if the sarcomas were ground and filtered to remove tumor cells, the filtrate induced serial sarcomas and thus the sarcoma was caused by a virus [123,124]. We now know that this is an XRV, Rous Sarcoma Virus (RSV), and the sarcoma is caused by the viral oncogene *v-src* [125–127]. Retroviruses can incorporate cellular proto-oncogenes into the viral genome which upon subsequent infection leads to the transformation of infected cells to support viral production. Other examples of acquired retroviral oncogenes include; *v-myc* in chicken myelocytomatosis virus [128]; *v-abl* in Abelson murine leukemia virus [129,130]; H- and K-*ras* in Harvey and Kirsten murine sarcoma virus, respectively [131,132]; *v-akt* in mouse AKT8 virus causing thymic lymphoma [133] and; *v-fms* (*CSF1R* in human) in McDonough feline sarcoma [134].

Retroviruses can also promote oncogenesis through insertional mutagenesis. Proviral DNA insertion sites are largely stochastic across the genome, with some viruses preferentially inserting into open chromatin. Yet in some cancers, such as avian leukosis virus (ALV)-induced T- and B-cell lymphomas, proviral DNA was recurrently found in the sense-orientation near the transcription start site (TSS) and first intron of the cellular *c-myc* gene [135]. The recurrent insertions are not the result of an insertion site preference of the virus, but it is the result of the selective advantage conferred by insertion at this location causing *c-Myc* over-expression. The cells with this particular insertion event undergo transformation and out-compete other uninfected cells or cells with

insertions in other, random locations. In this way, this is an oncogenic *driver* mutational event. Seminal examples of oncogenes identified by retroviral insertional mutagenesis include; *Wnt-1* by mouse mammary tumor virus [136,137] and *lck* and *c-Myc* mutation by Moloney murine leukemia virus causing T-cell lymphoma [138,139].

Insertional mutagenesis persists following retroviral endogenization and is a source of oncogenic mutation. In mice, ERV retrotransposition rates are high, responsible for ~10% of spontaneous phenotypic mutations [20]. As murine ERVs disperse across the genome they can over-express cellular proto-oncogenes like their exogenous cousins. Insertions of the murine intracisternal A-type particle ERV have led to myelomonocytic leukemia by causing *GM-CSF* over-expression and to T-cell lymphoma by inducing *IL3* overexpression [140]. In probably one of the most fascinating case studies, *three* independent loci of endogenous MLV in a immunodeficient *RAG1*$^{-/-}$ strain of mice, recombined to form an exogenous virus. The reconstituted MLV then transmitted horizontally to litter mates leading to collapse of the colony due to retroviral induced lymphoma, in two separate instances [15].

Human immunodeficiency virus (HIV), or any human retrovirus, has not been implicated in causing cancer through acquisition of a proto-oncogene or insertional mutagenesis. Although, HIV is associated with some cancers such as Kaposi Sarcoma, but this arises secondary to acquired immunodeficiency syndrome (AIDS) caused by HIV [141]. To date, most studies into potential roles for ERVs in human cancer have focused on their protein products. Indeed, there is strong evidence that the accessory proteins Np9 and Rec, encoded by members of the relatively young HERV-K (HML-2) group, have oncogenic properties, particularly in germ cell tumors [142–144]

Human T-lymphotropic Virus (HTLV) 1-4 are a family of retrovirus infecting humans. HTLV-1, which is the most prevalent of these viruses, infects between 10-20 million people worldwide [145]. HTLV-1 is also the only known oncogenic retrovirus of humans, causing a form of acute T-cell

leukemia [146]. HTLV-1 associated cell immortalization and transformation is mediated by the *tax* gene, encoded in the U3 region of the 3' LTR. The oncogenic capacity of HTLV-1 is mediated by the Tax protein, which interacts with CREB and p300/CBP to modulate cellular gene expression and ultimately leads to the inactivation of the tumour suppressor p53 [146].

Regardless of their retrotranspositional or coding capacity, ERVs may play a broader role in oncogenesis involving the intrinsic regulatory capacity of the LTR. De-repression/activation of cryptic (or normally dormant) promoters to drive ectopic expression is one mechanism by which the hundreds of thousands of dispersed ERV sequences can promote oncogenesis. I termed this distinct mechanism *onco-exaptation*.

1.2.3 Onco-exaptation of ERVs

The transcriptional up-regulation of LTR promoters and to a lesser extent L1 promoters are widespread in epigenetically perturbed cells such as cancer [147,148,117,149]. Here I discuss specific published examples of such onco-exaptation of TE promoters in affecting protein-coding genes (Figure 1.2). Although many TE-initiated transcripts have been identified [76], in this section I restrict the discussion to those cases where some role of the TE-driven gene in cancer or cell growth has been demonstrated.

1.2.3.1 Ectopic and overexpression of protein-coding genes

The most straightforward interaction between a TE promoter and a gene is when a TE promoter is activated, initiates transcription, and transcribes a downstream gene without altering the open reading frame (ORF), thus serving as an alternative promoter. Since the TE promoter may be regulated differently than the native promoter, this can result in ectopic and/or overexpression of the gene, with oncogenic consequences.

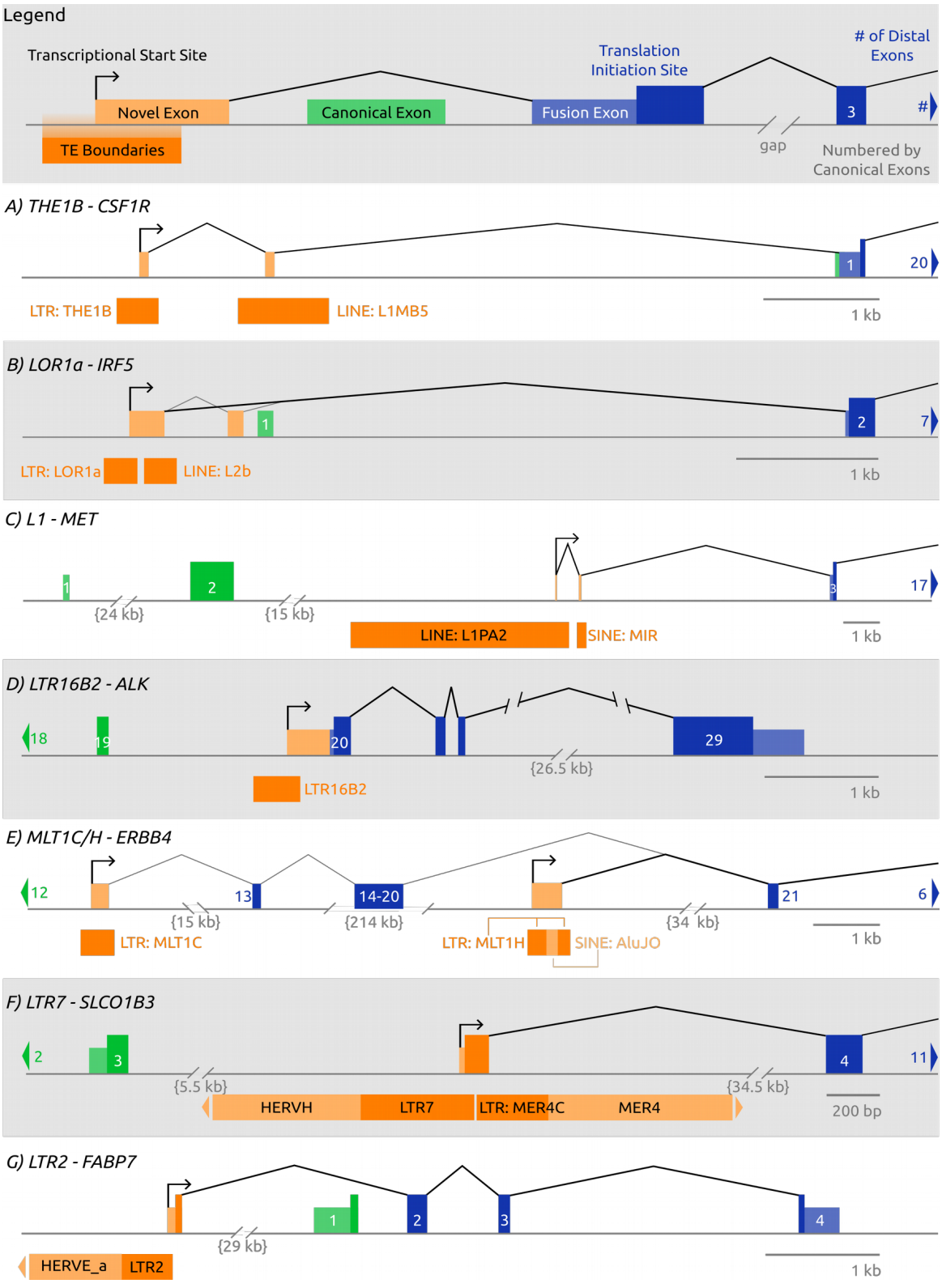


Figure 1.2: Examples of onco-exaptation

Text 1: Figure 1.2 Continued...

Gene models of known TE-derived promoters expressing downstream oncogenes. Legend is shown at the top. A) 6 kb upstream of CSF1R, a THE1B LTR initiates transcription and contains a splice donor site which joins to an exon within a LINE L1MB5 element and then into the first exon of CSF1R. The TE-initiated transcript has a different, longer 5' UTR than the canonical transcript but the same full-length protein coding sequence. B) An LOR1a LTR initiates transcription and splices into the canonical second exon of IRF5 that contains the standard translational initiation site (TIS) to produce a full-length protein. There also is a novel second exon which is non-TE derived which is incorporated into a minor isoform of LOR1a-IRF5 (see Chapter 4). C) Within the canonical intron 2 of the proto-oncogene MET, a full length LINE L1PA2 element initiates transcription (anti-sense to itself), splicing through a short exon in a SINE MIR element and into the third exon of MET. The first TIS of the canonical MET transcript is 14 bp into exon 2, although an alternative TIS exists in exon 3, which is believed to also be used by the L1-promoter 'd isoform'. D) An LTR16B2 element in intron 19 of the ALK gene initiates transcription and transcribes into the canonical exon 20 of ALK. An in-frame TIS within the 20th exon results in translation of a shortened oncogenic protein containing only the intra-cellular tyrosine kinase domain, but lacking the transmembrane and extracellular receptor domains of ALK. E) There are two TE-promoted isoforms of ERBB4, the minor variant initiates in an MLT1C LTR in the 12th intron and the major variant initiates in a MLT1H LTR in the 20th intron. Both isoforms produce a truncated protein, although the exact translation start sites are not defined. F) In the third exon of SLCO1B3, two adjacent partly full-length HERV elements conspire to create a novel first exon. Transcription initiates in the anti-sense orientation from an LTR7 and transcribes to a sense-oriented splice donor in an adjacent MER4C LTR, which then splices into the fourth exon of SLCO1B3, creating a smaller protein. G) An LTR2 element initiates anti-sense transcription (relative to its own orientation) and splices into the native second exon of FABP7. The LTR-derived isoform has a non-TE TIS and splice donor which creates a different N-terminal protein sequence of FABP7.

The first case of such a phenomenon was discovered in the investigation of a potent oncogene colony stimulating factor one receptor (*CSF1R*, also called *c-fms)* in Hodgkin Lymphoma (HL). Normally, *CSF1R* expression is restricted to macrophages in the myeloid lineage. To understand how this gene is expressed in HL, a B-cell derived cancer, Lamprecht et al. [150] performed 5` RACE which revealed that the native, myeloid-restricted promoter is silent in HL cell lines, with *CSF1R* expression instead being driven by a solitary THE1B LTR, of the MaLR-ERVL class (Figure 1.2A). THE1B LTRs are ancient, found in both Old and New World primates, and are highly abundant in the human genome, with a copy number of ~17,000 [2,151]. The *THE1B-CSF1R* transcript produces a full-length protein in HL, which is required for growth/survival of HL

cell lines [150] and is clinically prognostic for poorer patient survival [152]. Ectopic *CSF1R* expression in HL appears to be completely dependent on the THE1B LTR, and CSF1R protein or mRNA is detected in 39-48% of HL patient samples [150,152,153]. Another example of this type involving the IRF5 gene (Figure 1.2B) which was uncovered in my work and will be discussed in Chapter 4.

1.2.3.2 Expression of truncated proteins

In these cases, a TE-initiated transcript results in the expression of a truncated ORF of the affected gene, typically because the TE is located in an intron, downstream of the canonical translational initiation site. The TE initiates transcription, but the final transcript structure depends on the position of downstream splice sites, and protein expression requires usage of a downstream ATG. Protein truncations can result in oncogenic effects due to loss of regulatory domains or through other mechanisms, with a classic example being *v-myb*, a truncated form of *myb* carried by acutely transforming animal retroviruses [154,155].

The first such reported case involving a TE was identified in a screen of human ESTs to detect transcripts driven by the antisense promoter within L1 elements. Mätlik et al. identified an L1PA2 within the second intron of the proto-oncogene *MET* (*MET* proto-oncogene, receptor tyrosine kinase) that initiates a transcript by splicing into downstream *MET* exons [156] (Figure 1.2C). Not surprisingly, transcriptional activity of the CpG rich promoter of this L1 in bladder and colon cancer cell lines is inversely correlated to its degree of methylation [157,158]. A truncated MET protein is produced by the TE-initiated transcript and one study reported that L1-driven transcription of MET reduces overall MET protein levels and receptor signaling, although by what mechanism is not clear [158]. Analyses of normal colon tissues and matched primary colon cancers and liver metastatic samples showed this L1 is progressively demethylated in the metastasis samples, which correlates

with increased L1-MET transcripts and protein levels [159]. Since MET levels are a negative prognostic indicator for colon cancer [160], these findings suggest an oncogenic role for L1-MET.

More recently, Wiesner *et al.* identified a novel isoform of the receptor tyrosine kinase (RTK), anaplastic lymphoma kinase (*ALK)*, initiating from an alternative promoter in its 19th intron [161] (Figure 1.2D). This alternative transcription initiation (ATI) isoform or *ALK*^{ATI} was reported to be specific to cancer samples and found in ~11% of skin cutaneous melanomas. *ALK*^{ATI} transcripts produce three protein isoforms encoded by exons 20 to 29. These smaller isoforms exclude the extracellular domain of the protein but contain the catalytic intracellular tyrosine kinase domain. In neuroblastoma, the absence of *ALK*^{ATI} is a positive prognostic marker, predicting 5-year patient survival [162]. This same region of *ALK* is commonly found fused with a range of other genes via chromosomal translocations in lymphomas and a variety of solid tumors [163]. In the Wiesner et al. study it was found that ALK^{ATI} stimulates several oncogenic signaling pathways, drives cell proliferation in vitro, and promotes tumor formation in mice [161].

The *ALK*^{ATI} promoter is a sense-oriented solitary LTR (termed LTR16B2) derived from the ancient ERVL family. LTR16B2 elements are found in several hundred copies (Table 1.2) in the genomes of both primates and rodents [2,164] and this particular element is present in the orthologous position in mouse. Therefore, the promoter potential of this LTR has been retained for at least 70 Ma. Although not the first such case, the authors state that their findings “suggest a novel mechanism of oncogene activation in cancer through *de novo* alternative transcript initiation”. Evidence that this LTR is at least occasionally active in normal human cells comes from Capped Analysis of Gene Expression (CAGE) analysis through the FANTOM5 project [165]. A peak of CAGE tags from monocyte-derived macrophages and endothelial progenitor cells occurs within this

LTR, 60 bp downstream of the TSS region identified by Wiesner *et al*. [161], although a biological function of this isoform in normal cells is unknown.

To gain a molecular understanding of ALK-negative anaplastic large-cell lymphoma (ALCL) cases, Scarfo *et al.* conducted gene expression outlier analysis and identified high ectopic co-expression of *ERBB4* and *COL29A1* in 24% of ALCL cases [166]. Erb-b2 receptor tyrosine kinase 4 (*ERBB4*), also termed HER4, is a member of the ERBB family of RTKs, which includes EGFR and HER2, and overexpression of this gene have been implicated in some cancers [167]. Analysis of the ERRB4 transcripts expressed in these ALCL samples revealed two isoforms initiated from alternative promoters, one within intron 12 (I12-ERBB4) and one within intron 20 (I20-ERBB4), with little or no expression from the native/canonical promoter. Both isoforms produce truncated proteins that show oncogenic potential, either alone (I12 isoform) or in combination. Remarkably, both promoters are LTR elements of the ancient MaLR-ERVL class (Figure 1.2E). Of note, Scarfo *et al.* reported that two thirds of ERBB4 positive cases showed a “Hodgkin-like” morphology, which is normally found in only 3% of ALCLs [166]. We therefore examined our RNA-seq data from 9 HL cell lines and B-cell controls (Chapter 4) and found evidence for transcription from the intron 20 MLTH2 LTR in two of these lines, suggesting that truncated ERBB4 may play a role in some HLs.

In a screen for recurrent and cancer-specific TE-initiated transcripts in colorectal carcinoma, our group identified a second intron MSTD LTR element driving the over-expression of a truncated *IL33.* CRC cell lines expressing the *MSTD-IL33*, had increased efficiency to form 3-D colonospheres *in vitro* relative to *IL33* knockdown controls, functionally implicating this isoform [168]. More recently, IL-33 has been implicated as tumor immunosuppressant, through activation of effector T regulatory cells [169], but it remains to be determined if *MSTD-IL33* is capable of a similar function.

1.2.3.3 TE-promoted expression of chimeric proteins

Perhaps the most fascinating examples of onco-exaptation involve generation of a novel "chimeric" ORF via usage of a TE promoter that fuses otherwise non-coding DNA to downstream gene exons. These cases involve both protein and transcriptional innovation and the resulting product can acquire *de novo* oncogenic potential.

The solute carrier organic anion transporter family member 1B3, encoding organic anion transporting polypeptide 1B3 (*OATP1B3*, or *SLCO1B3*), is a 12-transmembrane transporter with normal expression and function restricted to the liver [170]. Several studies have shown that this gene is ectopically expressed in solid tumors of non-hepatic origin, particularly colon cancer [170–173]. Investigations into the cause of this ectopic expression revealed that the normal liver-restricted promoter is silent in these cancers, with expression of "cancer-type" (Ct)-OATP1B3 being driven from an alternative promoter in the second canonical intron [172,173]. While not previously reported as being within a TE, Lock et. al noted that this alternative promoter maps within the 5' LTR (LTR7) of a partly full-length antisense HERV-H element that is missing the 3' LTR [168]. Expression of HERV-H itself and LTR7-driven chimeric long non-coding RNAs is a noted feature of embryonic stem cells and normal early embryogenesis, where several studies indicate an intriguing role for this ERV group in pluripotency (for recent reviews see [81,174,175]). A few studies have also noted higher general levels of HERV-H transcription in colon cancer [176,177]. The LTR7-driven isoform of *SLCO1B3* makes a truncated protein lacking the first 28 amino acids but also includes protein sequence from the LTR7 and an adjacent MER4C LTR (Figure 1.2F). The novel protein is believed to be intracellular and its role in cancer remains unclear. However, one study showed that high expression of this isoform is correlated with reduced progression-free survival in colon cancer [178].

In another study designed specifically to look for TE-initiated chimeric transcripts, our laboratory screened RNA-seq libraries from 101 patients with diffuse large B-cell lymphoma (DLBCL) of different subtypes [179] and compared to transcriptomes from normal B-cells. This screen resulted in the detection of 98 such transcripts that were found in at least two DLBCL cases and no normals [180]. One of these involved the gene for fatty acid binding protein 7 (*FABP7*). FABP7, normally expressed in brain, is a member of the FABP family of lipid chaperons involved in fatty acid uptake and trafficking [181]. Overexpression of *FABP7* has been reported in several solid tumor types and is associated with poorer prognosis in aggressive breast cancer [181,182]. In 5% of DLBCL cases screened, Lock et al., found that *FABP7* is expressed from an antisense LTR2 (the 5' LTR of a HERV-E element) (Figure 1.2G). Since the canonical ATG is in the first exon of *FABP7*, the LTR driven transcript encodes a chimeric protein with a different N-terminus (see accession NM_001319042.1) [180]. Functional analysis in DLBCL cell lines revealed that the LTR-FABP7 protein isoform is required for optimal cell growth and also has sub-cellular localization properties distinct from the native form [180].

Overall, among all TE types giving rise to chimeric transcripts detected in DLBCL, LTRs were over represented compared to their genomic abundance and, among LTR groups, our group found that LTR2 elements and THE1 LTRs were over represented [180]. As discussed above, this predominance of LTRs over other TE types is expected.

Finally, a recent study revealed that, in hepatocellular carcinoma (HCC), an AluJb element upstream of the oncogene *LIN28B* can act as an alternative promoter generating a *LIN28B*-tumour-specific transcript (TST). The *LIN28B-TST* contains additional N-terminal amino acids relative to the wildtype LIN28B. The presence of the *LIN28B-TST* was negatively correlated with patient survival and shown to influence cell proliferation *in vitro* [183].

1.2.4 TE-initiated non-coding RNAs in cancer

Since TEs, particularly ERV LTRs, provide a major class of promoters for long non-coding RNAs [76,77,184], it is not surprising that multiple LTR-driven lncRNAs have been shown to be involved in cancer. These cases can be broadly divided into those with direct, measurable oncogenic properties and those with expression correlated with a cancer. Unlike the coding genes discussed above that have non-TE or native promoters in normal tissues, these lncRNAs are typically LTR-driven in normal or malignant cells.

1.2.4.1 TE-initiated lncRNAs with oncogenic properties

In an extensive study, Presner et al. reported that the lncRNA SchLAP1 (SWI/SNF complex antagonist associated with prostate cancer 1) is overexpressed in ~25% of prostate cancers, is an independent predictor of poor clinical outcomes and is critical for invasiveness and metastasis [185]. They found that *SchLAP1* inhibits the function of the SWI/SNF complex, which is known to have a tumor suppressor roles [186]. While not mentioned in the main text, the authors report in supplementary data that the promoter for this lncRNA is an LTR (Figure 1.3A). Indeed, this LTR is a sense-oriented solitary LTR12C (of the ERV9 group).

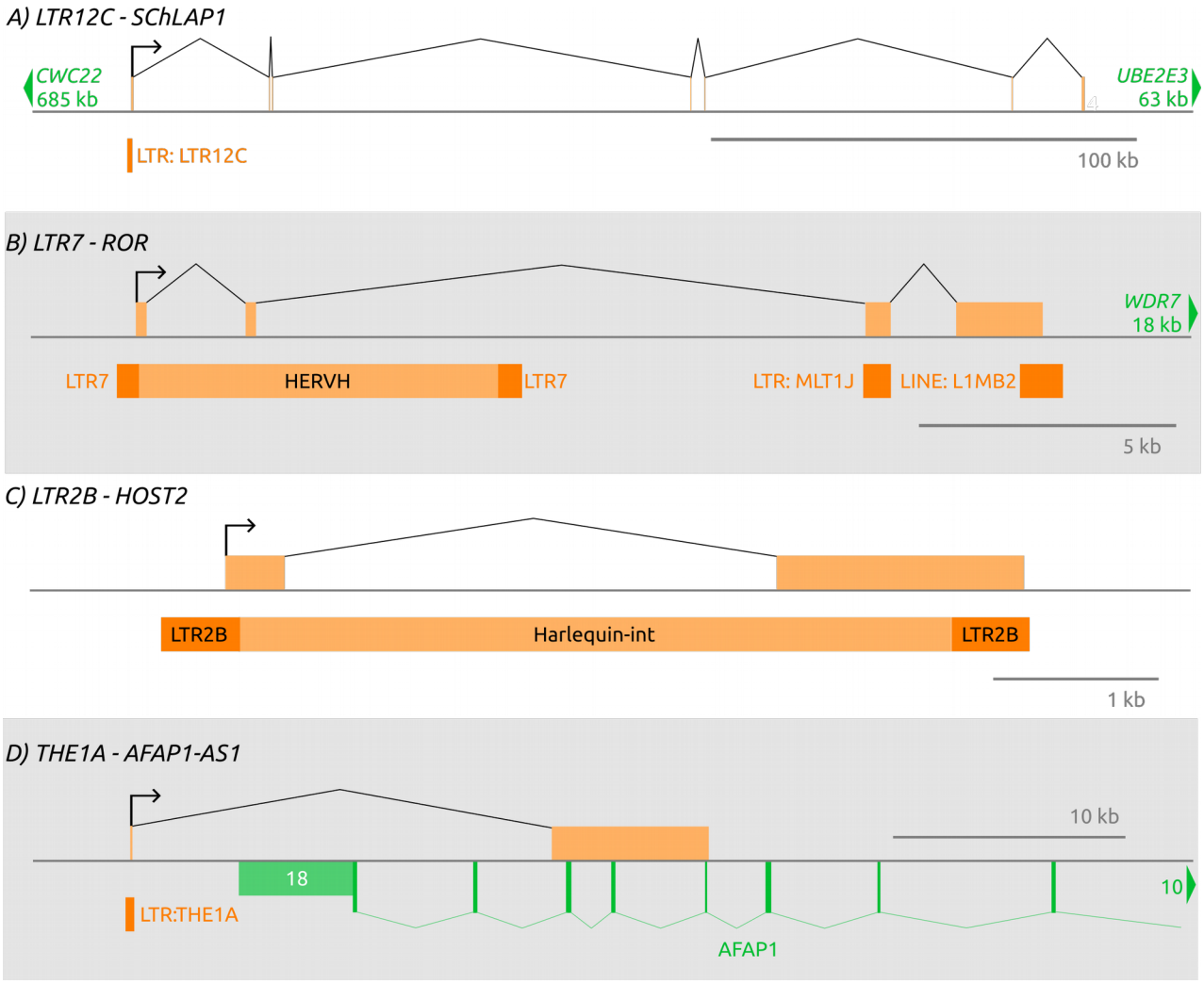


Figure 1.3: Examples of TE-initiated non-coding RNAs

Gene models of select lncRNAs initiating within LTRs that are involved in oncogenesis. ***A)*** *A solitary LTR12C element initiates SChLAP1, a long inter-genic non-coding RNA.* ***B)*** *The 5' LTR7 of a full-length HERVH element initiates the lncRNA ROR, with an exon partially incorporating internal ERV sequence.* ***C)*** *The HOST2 lncRNA is completely derived from components of a Harlequin (or HERV-E) endogenous retrovirus and its flanking LTR2B.* ***D)*** *Anti-sense to the AFAP1 gene, a THE1A LTR initiates transcription of the lncRNA AFAP1-AS1. The second exon of AFAP1-AS1 overlaps exons 14-16 of AFAP1, possibly leading to RNA interference of the gene.*

Linc-ROR is a non-coding RNA (long intergenic non-protein coding RNA, regulator of reprogramming) promoted by the 5' LTR (LTR7) of a full length HERV-H element [77] (Figure

1.3B) and has been shown to play a role in human pluripotency [187]. Evidence suggests it acts as a microRNA sponge of miR-145, which is a repressor of the core pluripotency transcription factors Oct4, Nanog and Sox2 [188]. Several recent studies have reported an oncogenic role for *Linc-ROR* in different cancers by sponging miR-145 [189–191] or through other mechanisms [192,193].

Using Serial Analysis of Gene Expression (SAGE), Rangel *et al.* identified five Human Ovarian cancer Specific Transcripts (HOSTs) that were expressed in ovarian cancer but not in other normal cells or cancer types examined [194]. One of these, *HOST2*, is annotated as a spliced lncRNA entirely contained within a full length HERV-E and promoted by an LTR2B element (Figure 1.3C). My perusal of RNA-Seq from the 9 core ENCODE cell lines shows robust expression of *HOST2* in GM12878, a B-lymphoblastoid cell line, which extends beyond the HERV-E. As with *Linc-ROR*, *HOST2* appears to play an oncogenic role by functioning as a miRNA sponge of miRNA *let-7b*, an established tumor suppressor [195], in epithelial ovarian cancer [196].

The lncRNA *AFAP1 antisense RNA 1* (*AFAP1-AS1*) runs antisense to the actin filament associated protein 1 (*AFAP1*) gene and several publications report its up-regulation and association with poor survival in a number of solid tumor types [197–200]. While the oncogenic mechanism of *AFAP1-AS1* has not been extensively studied, one report presented evidence that it promotes cell proliferation by upregulating RhoA/Rac2 signaling [201] and its expression inversely correlates with *AFAP1*. Although clearly annotated as initiating within a solitary THE1A LTR (Figure 1.3D), this fact has not been mentioned in previous publications. In screens for TE-initiated transcripts using RNA-seq data from HL cell lines, I noted recurrent and cancer-specific up-regulation of *AFAP1-AS1* (unpublished observations), suggesting that it is not restricted to solid tumors. The inverse correlation of expression between AFAP1 and AFAP1-AS1 suggests an interesting potential mechanism by which TE-initiated transcription may suppress a gene; where an anti-sense TE-

initiated transcript disrupts the transcription, translation or stability of a tumor suppressor gene transcript through RNA interference [202].

The *SAMMSON* lncRNA (survival associated mitochondrial melanoma specific oncogenic non-coding RNA), which is promoted by a solitary LTR1A2 element, was recently reported as playing an oncogenic role in melanoma [203]. This lncRNA is located near the melanoma-specific oncogene *MITF* and is always included in genomic amplifications involving *MITF*. Even in melanomas with no genomic amplification of this locus, *SAMMSON* is expressed in most cases, increases growth and invasiveness and is a target for SOX10 [203], a key TF in melanocyte development which is deregulated in melanoma [204]. Interestingly, the two SOX10 binding sites near the SAMMSON TSS lie just upstream and downstream of the LTR, suggesting that both the core promoter motifs provided by the LTR and adjacent enhancer sites combine to regulate *SAMMSON* [205].

Other examples of LTR-promoted oncogenic lncRNAs include *HULC* for Highly Upregulated in Liver Cancer [206,207], *UCA1* (urothelial cancer associated 1) [208–210] and *BANCR* (BRAF-regulated lncRNA 1) [211–213]. Although not mentioned in the original paper, three of the four exons of *BANCR* were shown to be derived from a partly full length MER41 ERV, with the promoter within the 5’LTR of this element annotated MER41B [76]. Intriguingly, MER41 LTRs were recently shown to harbor enhancers responsive to interferon, indicating a role for this ERV group in shaping the innate immune response in primates [104]. It would be interesting to investigate roles for *BANCR* with this in mind.

1.2.4.2 TE-initiated lncRNAs as cancer-specific markers

There are many examples of TE-initiated RNAs with potential roles in cancer or which are preferentially expressed in malignant cells but for which a direct oncogenic function has not yet been demonstrated. Still, such transcripts may underlie a predisposition for transcription of specific

groups of LTRs/TEs in particular malignancies and therefore function as a marker for a cancer or cancer subtype. Since these events potentially do not confer a fitness advantage for the cancer cell, they are not “exaptations” but “nonaptations” [7].

One of these is a very long RNA initiated by the antisense promoter of an L1PA2 element as reported by Tufarelli’s group and termed *LCT13 [214,215]*. EST evidence indicates splicing from the L1 promoter to the *GNTG1* gene, located over 300 kb away. The tumor suppressor gene, tissue factor pathway inhibitor 2, **(***TFPI-2)*, which is often epigenetically silenced in cancers [216], is antisense to LCT13 and it was shown that LCT13 transcript levels are correlated with down regulation of *TFPI-2* and associated with repressive chromatin marks at the *TFPI-2* promoter [215].

Gibb *et al.* analyzed RNA-Seq from colon cancers and matched normal colon to find cancer-associated lncRNAs and identified an RNA promoted by a solitary MER48 LTR, which they termed *EVADR*, for Endogenous retroviral associated ADenocarcinoma RNA [148]. Screening of data from The Cancer Genome Atlas (TCGA) [217] showed that *EVADR* is highly expressed in several types of adenocarcinomas, it is not associated with global activation of MER48 LTRs across the genome and its expression correlated with poorer survival [148]. In another study, Gosenca *et al.* used a custom microarray to measure overall expression of several HERV groups in urothelial carcinoma compared to normal urothelial tissue and generally found no difference [218]. However, they found one full-length HERV-E element, located in the antisense direction in an intron of the *PLA2G4A* gene that is transcribed in urothelial carcinoma and appears to modulate *PLA2G4A* expression, thereby possibly contributing to carcinogenesis, although the mechanism is not clear.

By mining long nuclear RNA data-sets from ENCODE cell lines, normal blood and Ewing sarcomas, one group identified over 2000 very long (~50-700 kb) non coding transcripts termed vlincRNAs [184]. They found the promoters for these vlincRNAs to be enriched in LTRs, particularly for cell type-specific vlincRNAs, and the most common transcribed LTR types varied in

different cell types. Moreover, among the data-sets examined, they reported that the number of LTR-promoted vlincRNAs correlated with degree of malignant transformation, prompting the conclusion that LTR-controlled vlincRNAs are a “hallmark” of cancer [184].

In a genome-wide CAGE analysis of 50 hepatocellular carcinoma (HCC) primary samples and matched non-tumor tissue, Hashimoto *et al*. found that many LTR-promoted transcripts are upregulated in HCC, most of these apparently associated with non-coding RNAs as the CAGE peaks in the LTRs are far from annotated protein coding genes [147]. Similar results were found in mouse HCC. Among the hundreds of human LTR groups, they found the LTR-associated CAGE peaks to be significantly enriched in LTR12C (HERV9) LTRs and mapped the common TSS site within these elements, which agrees with older studies on TSS mapping of this ERV group [219]. Moreover, this group reported that HCCs with highest LTR activity mostly had a viral (Hepatitis B) etiology, were less differentiated and had higher risk of recurrence [147]. This study suggests widespread tissue-inappropriate transcriptional activity of LTRs in HCC.

1.3 Thesis objectives

Transposable element transcriptional initiation has been associated with different epigenetic perturbations, yet previous studies have been dependent on specialized assays, either focusing on a sub-set of TEs, or on initiation sites in the absence of their transcriptomic consequences. This thesis creates a generalizable platform for TE-initiation detection with simultaneous inference on the transcriptional consequences. Together this allowed for a detailed analysis of TE transcription, ultimately exploring novel applications for TE-initiated transcription.

Thesis Hypothesis: Cancer transcriptomes have increased transposable element transcription relative to normal cell of origin controls.

Corollary: Increased transposable element transcription accelerates tumorigenesis.

The objective of Chapter 2 was to develop a bioinformatic tool to detect and quantify TE-derived promoters genome wide. This was accomplished by the transcriptome sequencing analysis suite *LIONS*, that outputs an annotation of TE-initiated transcripts per sequencing library. The outcome of this work was that TE-initiated transcripts can be globally quantified, grouped and compared across biological groups from RNA-seq data alone.

The objective of Chapter 3 was to measure the global contribution of TEs in cancer and normal transcriptomes, and in response to cellular state changes associated with epigenetic perturbation. This was accomplished by applying *LIONS* to colorectal carcinoma and patient-matched normal RNA-seq, as well as two models for cellular senescence. The outcome of these analyses was the global characterization of TE de-repression in cancer and senescence and an analysis of the underlying distributions that ultimately control TE-initiated transcription.

The objective of Chapter 4 was an in-depth analysis of the biological consequences of TE-initiated transcription in Hodgkin lymphoma and diffuse large B-cell lymphoma. This includes a

case study of the LTR onco-exaptation of *IRF5* and exploring the use of TE-initiated transcription as a diagnostic biomarker.

The objective of the final Chapter 5 was to conclude with a theoretical model with which the data in the previous chapters can be interpreted. This was accomplished through the synthesis of the data and the literature. This model may be useful as the basis with which future research on TE and ERV activity in cancer can be interpreted and be employed to develop novel prognostic technologies for the benefit of human cancer patients.

Chapter 2: *LIONS:* Detection and quantification of transposable element derived promoters in RNA-seq

2.1 Background

The percentage of transcripts initiated within repetitive DNA as measured by Cap Analysis Gene Expression (CAGE) is substantial, ranging from ~3-15% in humans depending on the tissue [75]. Such TE-initiated transcripts are enriched for long non-coding RNAs (lncRNA) [76,77]. In human embryonic stem cells (hESCs), ERV transcription in particular is a marker of pluripotency, as it is in mice [220]. There is also growing evidence that ERV-initiated transcripts are functionally involved in the evolution of the human pluripotent stem cell transcriptome [81,221–223].

TEs in the vicinity of protein coding genes may gain function over evolutionary time as alternative tissue-specific promoters, like the THE1D LTR element that drives placental-specific transcription of human *IL2RB* [101]. Interestingly, over the course of cancer evolution, normally dormant TE promoters can be exploited to express a protooncogene. Such “onco-exaptations” have been identified for the expression of *CSF1R* [150] and *IRF5* (Chapter 4, [224]) in Hodgkin Lymphoma, *FABP7* [180] in Diffuse Large B-cell Lymphoma and *ALK* in melanoma [161] among others (Chapter 1, [205]). While a number of cases of onco-exaptations have been documented, the mechanisms underlying these oncogenic events remains largely unexplored.

It has been proposed that TE invasions may function as evolutionary accelerants, promoting adaptation and correlating with the radiation of species [225,58] and therefore there is a significant interest in understanding the extent and evolutionary mechanisms by which TEs contribute to a cell's transcriptome. Previous transcriptome-wide studies designed to detect TE-derived promoters have analyzed annotated mRNAs [226], ESTs [227], assembled transcripts [77,76,228], short Cap Analysis Gene Expression CAGE tags [75], Paired-end ditag sequences [229], paired-end 'chimeric

fragment' RNA-seq screening [180,230,231], targeted TE events such as ERV9-driven [232] or L1-driven transcripts [215] and loci-gene correlation studies [233]. While these methods have proved useful, they have significant limitations.

5` CAGE is the clearest measure of transcription start sites (TSSs) but provides insufficient information on the resultant transcript structure. RNA-seq assembly methods may not identify the true 5' end of transcripts or suffer from a high false positive rate due to TE exonization events. The TE-exonization problem also creates high false-positive rates in chimeric fragment-based and hybridization-based methods that have gone unaddressed [230–232,234]. Moreover, none of the aforementioned studies have attempted to quantify the strength or contribution of the putative TE-initiated isoforms to overall transcript expression when alternative promoters exist. Therefore, effective TE-initiating transcript screens have required extensive human-inspection and have failed to provide a quantitative, genome-wide assessment of TEs initiating biologically significant transcription.

While there are many software packages to analyze TE mobilization at the DNA level or look at TE expression alone, there is no analysis software to quantify TE-initiation events from RNA-seq data [235,236]. To quantitatively measure and compare the contribution of TE promoters to normal and cancer transcriptomes I developed a tool that incorporates features of previous methods but significantly builds upon them. I was motivated to use paired-end RNA-seq data alone, a broadly available data-type, to rapidly measure TE-initiations and transcriptome contributions. With a defined set of TE-initiated transcripts in each library, commonalities and differences between sets of data (biological replicates) can be determined. Together these analyses have been packaged to give rise to the *LIONS* suite (Figure 2.1).

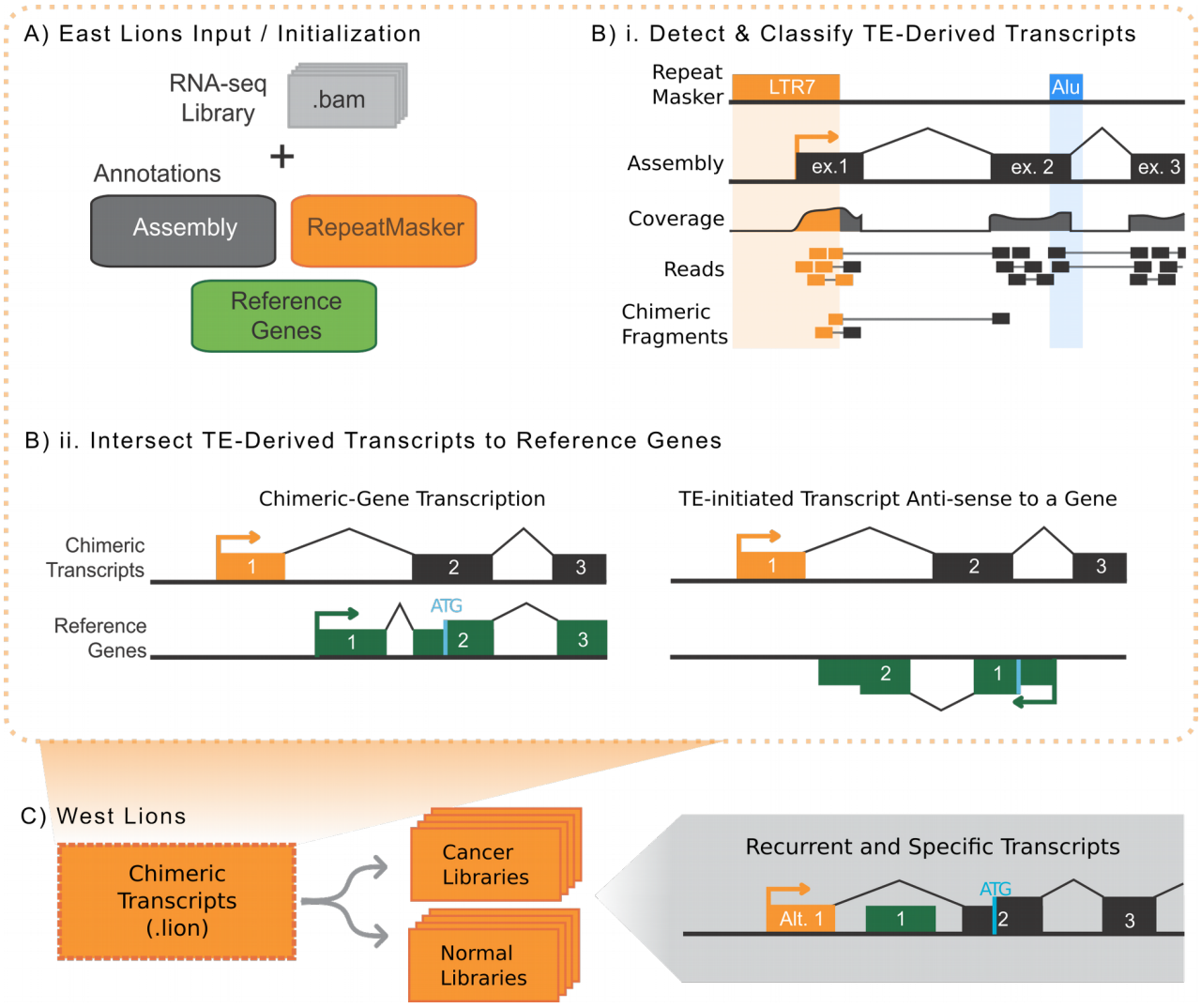


Figure 2.1: Schematic of LIONS workflow

The workflow for *LIONS* is divided into two main components. **A)** 'East Lion' analyzes individual transcriptomes starting with i) a .bam file(s) of paired-end reads, a reference genome, a RepeatMasker annotation and a reference set of protein coding genes. The reads are aligned to the genome with the spliced read mapper Tophat2 [237] and an *ab initio* transcriptome is assembled with Cufflinks [238]. ***B)*** I) These data are then analyzed per chimeric fragment cluster for transposable element (TE)-initiated transcripts (Figure 2.2A). Briefly, fragment clusters consistent with transcriptional initiation (Orange) are enriched and those with passive exonization (Blue) or termination (not shown) are depleted. ii) The set of TE-initiated contigs are then intersected to reference set of protein coding genes and classified with respect to their intersection. Each transcriptome is analyzed independently and a standard .lions output file is generated. **C)** 'West Lion' performs set analysis on the .lions files. Transcriptomes are biologically grouped and analyzed individually and as part of a biological group (i.e. cancer vs. normal samples).

2.2 Materials & methods

2.2.1 Initialization, alignment and assembly

For an accurate measurement of TE initiated transcripts starting from whole transcriptome sequencing data the *LIONS* software suite containing the *East Lion* and *West Lion* modules was developed (Figure 2.1). The central principle in detecting transcription start sites within TEs is that a local analysis is performed for patterns of sequencing reads consistent with transcriptional TE-initiation.

The primary *LIONS* input is a set of paired-end RNA sequencing data either in fastq or bam format. The data-sets can be biologically or technically grouped for later comparisons or individual libraries can be run. Additionally, a reference genome (hg19), a RepeatMasker [12] analysis of that genome (hg19 – 2009-04-24), and a set of reference protein-coding genes (UCSC Genes, 2013-06-14) is required. Reference annotations were up to date at the time this project was initiated.

A workspace for the project is initialized on the system and an optional alignment is run with the splice-aware aligner *tophat2* (v.2.0.13) [237] such that secondary alignments for multi-mapping reads are retained and flagged; *tophat2 --report-secondary-alignments.* On systems that support *qsub* parallelization and multiple CPU cores, each library is aligned in parallel with multiple threading allowing for rapid analysis of large data-sets.

Following alignment, *ab initio* transcriptome assembly is performed on each library using repeat-optimized parameters of *Cufflinks* (v.2.2.1) [238]; *cufflinks --min-frags-per-transfrag 10 --max-multiread-fraction 0.99 --trim-3-avgcov-thresh 5 –trim-3-dropoff-frac=0.1 --overlap-radius 50.* The use of an assembly substantially reduced false-positive TE-initiation calls relative to using a reference gene set since only transcript isoforms that exist in the data are considered, although it is possible to forego this step and use a reference gene set. The generated alignment and assembly is

then processed to generate a bigwig coverage file for visualization and basic statistics for each exon and TE are calculated such as read-coverage and RPKM.

2.2.2 Detection and classification of TE-initiated transcripts

To search the sequencing data for potential TE-exon interactions, each TE-exon pair for which a chimeric fragment cluster exists are considered. Briefly, a chimeric fragment cluster is a set of reads where one read maps to a TE and its pair maps to an exon from the assembly (Figure 2.2). These TE-exon pairs form the basis for classification into one of three cases; TE-initiation, -exonization or -termination of the transcript (Figure 2.2).

Classification is accomplished by the calculation of a series of values that are then fed into a classification algorithm. First, the relative position of the TE and exon boundaries with respect to the direction of transcription is compared. Only intersection cases in which the TE is upstream of the exon and could initiate transcription are considered (Figure 2.3A). A thread ratio is then calculated, the ratio of read pairs in which one read maps outside of a TE in either the downstream or upstream direction. A high thread ratio distinguishes TE-initiation events from TE-exonizations, that is to say, if a TE initiates transcription then there should exist a strong bias towards the number of read-pairs downstream of the element (Figure 2.3B).

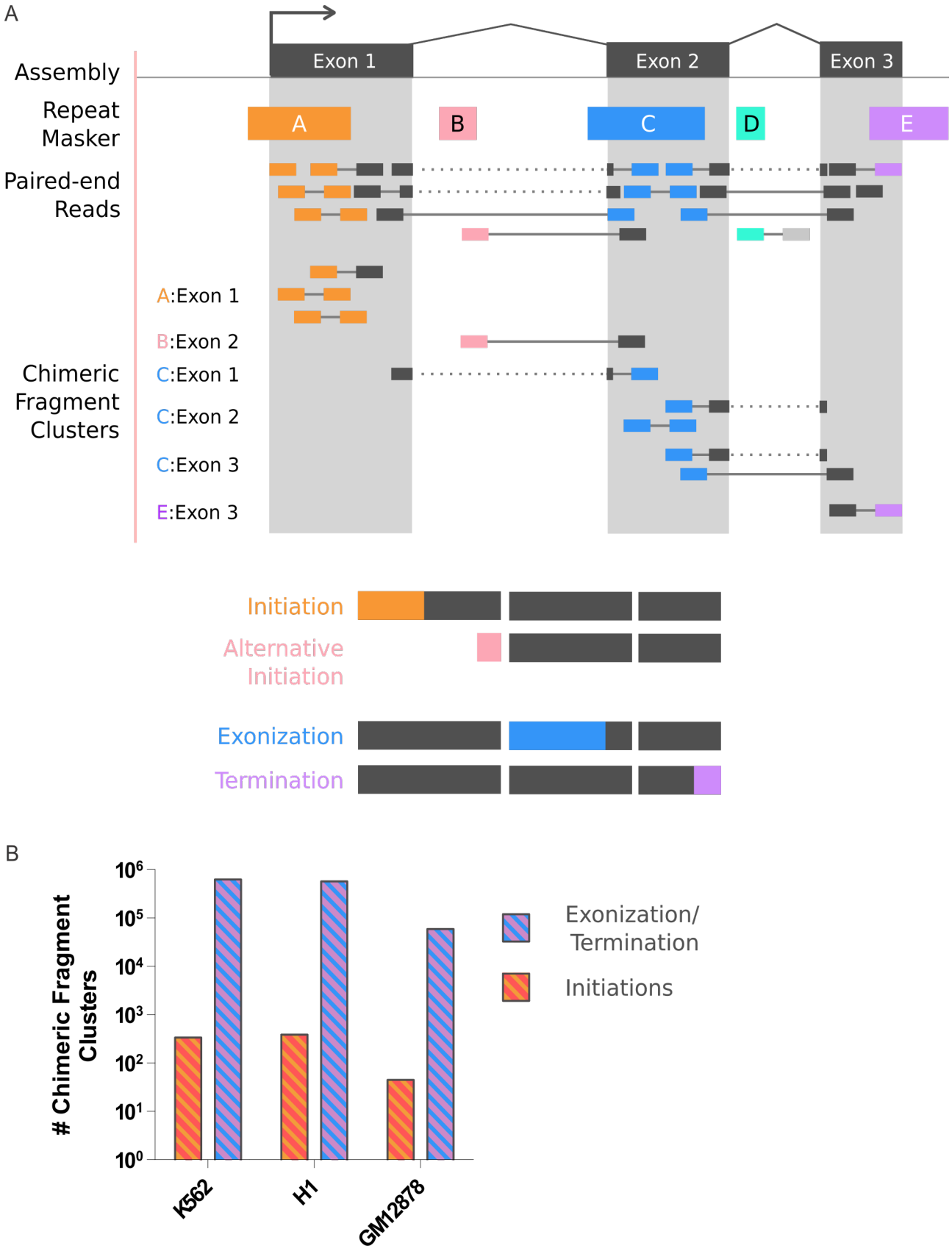


Figure 2.2: Chimeric fragment clustering in LIONS

Text 2: Figure 2.2: Continued.

A) *The analysis space of LIONS is all Repeat-Exon combinations for which there exists a chimeric fragment; paired-end sequencing reads in which one read intersects a repeat and the read pair intersects an exon from the assembly. Chimeric fragments can be yielded from an RNA molecule in three cases; i) TE-initiated transcripts (Repeat A:Exon 1 and Repeat B:Exon 2); ii) TE exonization in a transcript, either as a repeat is contained within an exon or the repeat is at an exon splice site (Repeat C:Exon 1,2 and 3) or iii) TE terminated transcripts (Repeat E:Exon 3). Each chimeric fragment cluster then is classified as either initiating a transcript or not based on local statistics for each repeat and exon pair such as; Repeat-Exon intersection, Exon and Repeat expression level, adjacent exon expression levels and read threading (Figure 2.3).* **B)** The number of chimeric fragments in K562, H1 or GM12878 transcripts that are classified as initiations compared to non-initiating clusters.

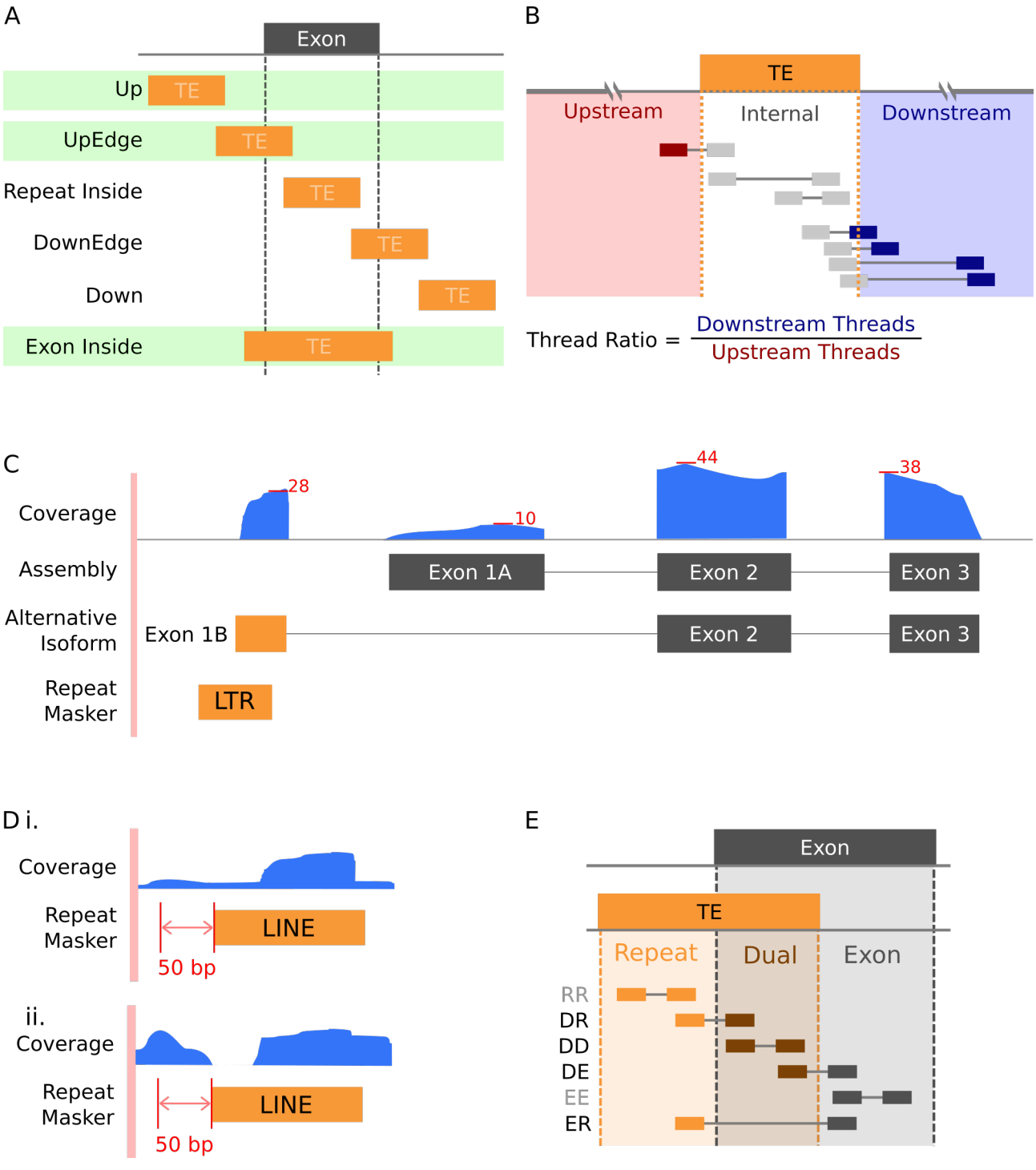


Figure 2.3: Calculated values for LIONS classification

To distinguish transposable element (TE)-initiated transcripts from TE exonizations or TE-terminated transcripts several local values are calculated for each chimeric fragment cluster. ***A)*** *The position of the TE (orange) relative to the exon (dark gray). Cases in which the TE is upstream, on the upstream edge, contained within the exon or contains the exon are considered for TE-initiation (highlighted green)...*

Text 3: Figure 2.3 Continued.

*… **B)** The thread ratio for a TE considers direction bias in sequencing read pairs going upstream or downstream relative to the interacting exon. Upstream threads (red) are read pairs in which one read maps to within the TE and the pair maps upstream of the TE. Downstream threads (blue) are the converse to upstream threads while read pairs with both reads internal to the TE are not counted (gray). The thread ratio is the number of downstream threads divided by the number of upstream threads, or set to the cut-off threshold when no upstream threads are present for inclusion. **C)** The contribution score is an approximation of the TE promoter contribution to the expression of downstream exons for alternative or unassembled TE promoter. The maximum coverage within the TE, 28 reads, is divided by the maximum coverage within the interacting exon (exon 2), 44 reads, to yield an approximate contribution for the TE-exon interaction, 0.636. **D)** The read coverage for the 50 bp immediately upstream of the TE is divided by the coverage of the TE itself to measure the background level of transcription at this loci. **i.** A locus with low levels of transcriptional readthrough but a potential initiation site present within the TE. **ii.** In contrast, a locus in which there is an apparent gain of coverage within the LINE but could be due to poor mapping quality at the 5` end of this LINE. **E)** Chimeric fragment sub-classification of whether a read intersects only a repeat (R), only an exon (E) or both (D). Chimeric fragments can thus be classified as DR, DD,DE or ER fragments. The ratio between the classifications can be used as a stringency cutoff for improving LIONS classification specificity. Taken together these values form the basis for LIONS classification of TE-initiated transcripts and are fed into the the sorting algorithms (Figure 2.4).*

For the detection of TE-initiated transcripts of biological significance further restrictions are imposed. Single exon contigs are excluded from the analysis to reduce the false positive rate (retained introns, low abundance lncRNAs). To quickly discard rare TE-initiated isoforms when an alternative, highly expressed isoform exists, TE contribution was estimated as the peak coverage within the TE divided by the peak coverage of its interacting exon (Figure 2.3C). Together these values form the basis on which TE-initiation, -exonization or -termination can be distinguished.

Classification of TE-exon interactions is performed by the sorting algorithm that can be customized (Figure 2.4). The default set of parameters termed, 'o*ncoexaptation'* were manually defined by extensive manual inspection of the training ENCODE sequencing data and comparison with supporting ChIP-seq and CAGE data such as shown for the *FHAD1* test-case (Figure 2.5). The default parameters are trained to conservatively detect high-abundance isoforms of TE-initiated transcripts with a biologically plausible contribution to overall gene expression and cancer biology.

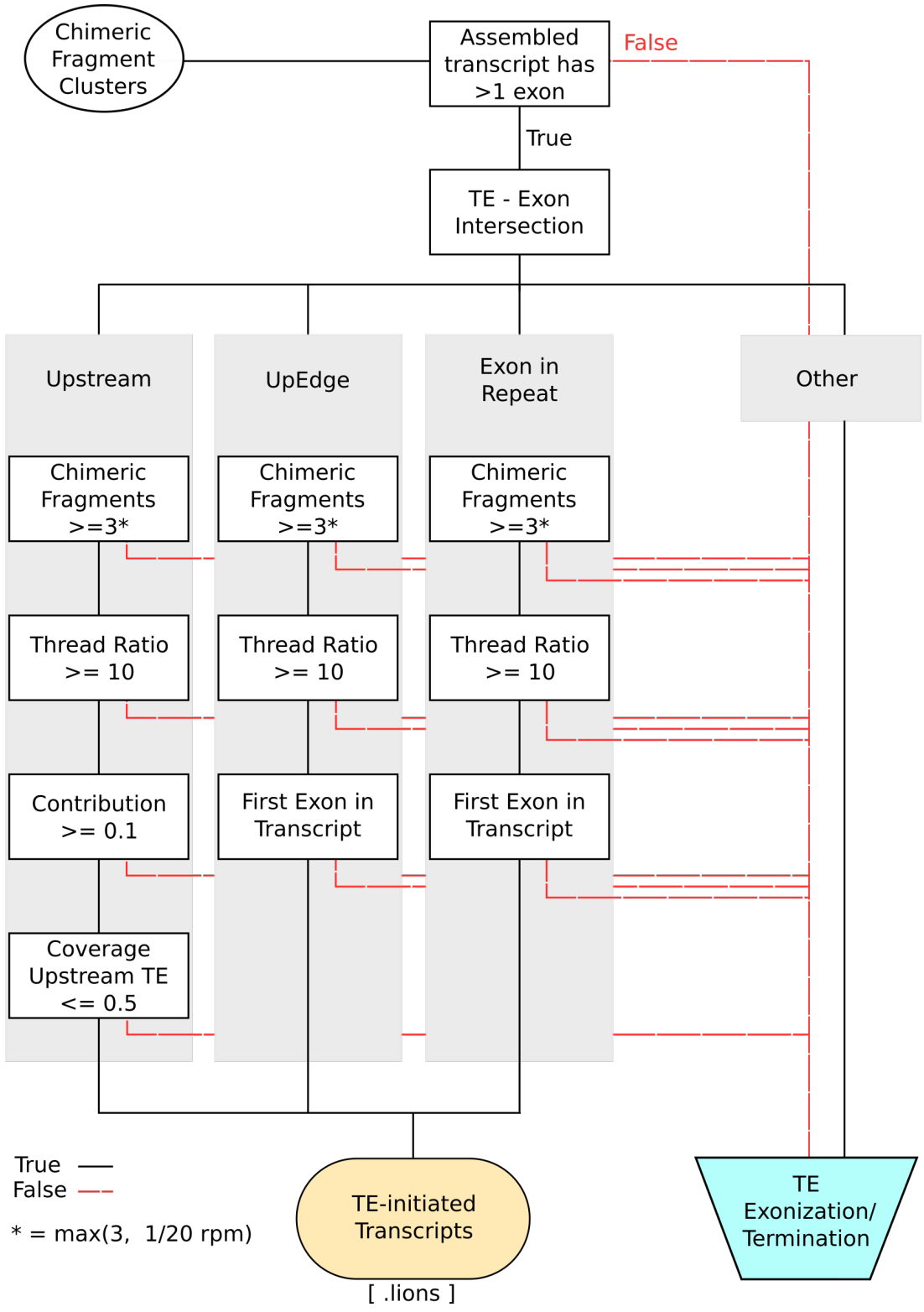


Figure 2.4: Chimeric fragment clusters sorting algorithm for TE-initiated transcripts

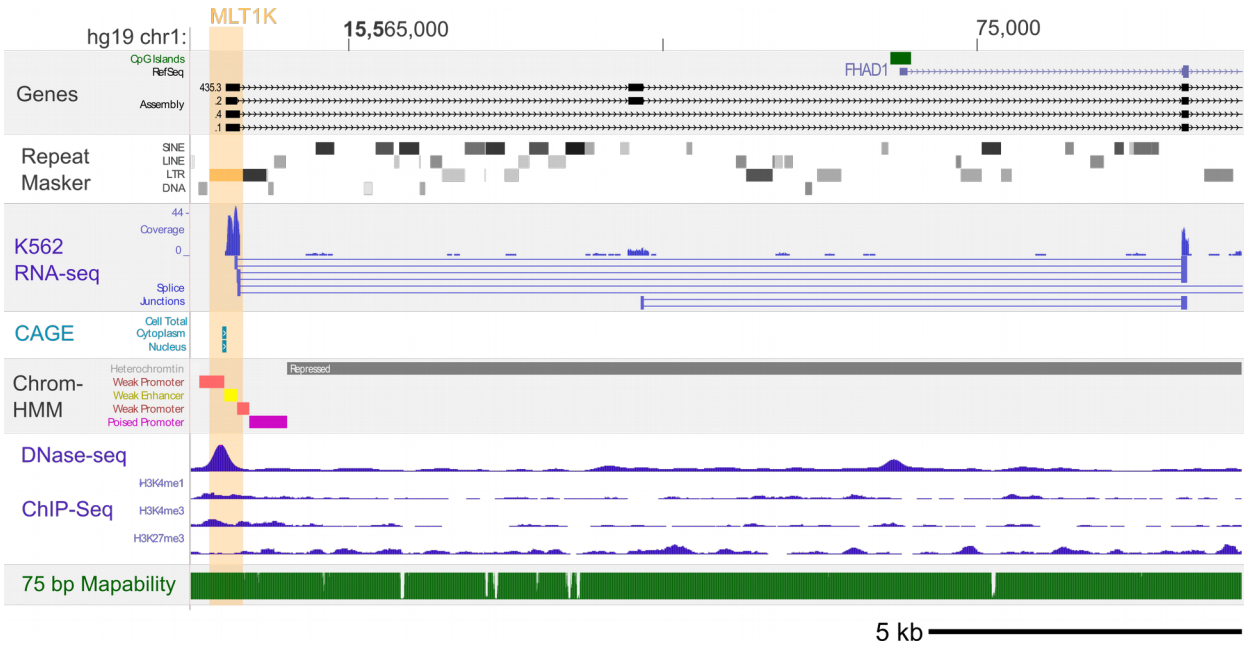


Figure 2.5: UCSC genome browser view of a LIONS identified chimeric transcript in K562

An upstream MLT1K LTR element initiates transcription and splices into exon 2 of the FHAD1 gene in which the coding sequence begins. The Cufflinks assembly contigs as well as the aligned reads and tophat2 detected splice junctions are shown, this case would be classified as 'Einside' as the entire first exon is contained within the MLT1K LTR element. CAGE hidden Markov model clusters (UCSC accessions: whole cell wgEncodeEH001150; cytosol wgEncodeEH000332; nucleus wgEncodeEH000333), DNase-seq (wgEncodeEH000530) and ChIP-seq (H3K4me1 wgEncodeEH000046; H3K4me3 wgEncodeEH000048; H3K27me3 wgEncodeEH000044) coverage support that this is a promoter as well as being classified as a 'weak promoter' by the respective Broad ChromHMM model (wgEncodeEH000790).

TE-initiated transcripts can be further sub-classified by their intersection to a set of protein-coding genes into; chimeric transcripts, TE-initiated transcripts that transcribe in the sense-orientation into a neighboring protein-coding gene; anti-sense TE-transcripts, non-coding TE-initiated transcripts which run anti-sense to a protein-coding gene; or long intergenic non-coding (linc) TE-transcripts which don't overlap a known protein-coding gene. Of particular interest to cancer biology are chimeric transcripts that result in the overexpression of oncogenes, such as previously identified in Hodgkin Lymphoma for *IRF5* and *CSF1R* [150,224].

Alternative algorithm filtering settings exist for algorithm parameters (in order: reads, thread ratio, downstream threads, exon RPKM, contribution score, upstream coverage and upstream exon RPKM) based on the experimental demand such as; '*screenTE'* (parameters: *2 5 5 1 0.05 2 1*), a sensitive but error-prone (exonizations called as initiations) method or; '*driverTE'* (parameters: *5 10 10 1 0.75 2 1.5)* detection of TE-initiated transcripts which are exclusively transcribed from TEs. Each of these settings are customizable and should be tailored towards individual project requirements. These analyses and filters are applied independent for each RNA-seq library and a standard .lion file is created. Sets of .lion files (that is sets of RNA-seq library analyses) are then grouped into a row-merged .lions file for set-based comparisons

2.2.3 Operating characteristics

To test the performance of the *LIONS* classification, a simulation of RNA-seq dataset as generated to benchmark the operating characteristics of the classifier. Starting with aligned RNA-seq from H1 hESCs and K562 chronic myeloid leukemia cell line, simulated transcriptomes were generated. For the first dataset, the top 20,000 expressed *gencode* transcripts in the K562 transcriptome, or in the second dataset the top 20,000 expressed assembled contigs from hESC transcriptome assembly were defined as the ‘reference transcriptome’ for simulation. *FluxSimulator [239]* was then used to simulated paired-end fastq based on these ‘reference transcriptomes’. From the K562 transcriptome, reads were simulated at 25, 100 and 200 million reads, yielding 14,610, 18,162, and 19,492 detectable TE-exon interactions, respectively. While the H1esc transcriptome was simulated at 5, 30, 100 and 200 millions reads, yielding 10,217, 16,781, 18,123, and 19,296 detectable TE-exon interactions, respectively. The simulated data were then processed by *LIONS* and compared to the input reference transcriptomes, which are defined as a ‘ground truth’ for this experiment.

2.2.4 Recurrent and group-specific TE-promoters

Grouping and comparing sets of TE-initiated transcripts is of central importance to understanding the biology of their activity. TE-initiated transcripts are more variable then non-TE transcripts across biological replicates (Figure 2.7) and therefore the TE signals from individual transcriptomes are noisy. The reasoning then is that grouping recurrent TE-initiated transcripts across biological replicates and asking which transcripts are recurrent will enrich for TE-initiated transcripts of consequence. In a similar line of reasoning, comparing one biological group against another can identify TE-initiated transcripts, or even classes of TEs that are more transcriptionally active in one group of transcriptomes relative to another.

To detect recurrent TE-initiated transcripts between libraries, the set of all TEs which initiate a transcript are considered (even if the downstream transcript structure is not the same). The recurrence cut-off parameter is the number of libraries within a test biological group that a given TE initiating transcription is required to be detected within. The specificity cut-off is the number of control libraries the initiating TE can also be detected in. Together, TEs which have greater than the recurrent cut-off and less than the specificity parameter cut-off are considered recurrent and specific TE-initiated transcripts for a test group (Figure 2.2).

A case in which recurrent and biological-group specific TE-initiated transcripts is significant is in cancer biology. The onco-exaptation hypothesis [205] predicts that the highly variable TE-initiated transcripts can be selected for during cancer evolution and therefore transcripts recurrent and cancer-specific are enriched for oncogenes or transcripts involved in the biology of the cancer.

2.2.5 RNA-seq data sets

ENCODE training RNA-seq fastq files were downloaded from the UCSC ENCODE ftp site. Hodgkin Lymphoma cell line and primary B-cell transcriptomes [179,224,240–242] bam files were converted to fastq for re-analysis by *LIONS*. Accession and library details are in Supplementary

Table 2.1. ENCODE data accessed at ftp://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeCaltechRnaSeq/ . Hodgkin Lymphoma cell culture, RNA isolation and cDNA synthesis was performed as described in Chapter 4 and [224]. Primers for RT-PCR are listed in Supplementary Table 2.2.

2.2.6 Brunswick: Artificial neural network classifier

An alternative classifier based on artificial neural network (ANN) was developed, called the *Brunswick* module. Simulated RNA-seq simulation data was used to train and test the operating characteristics of the ANN. The simulated RNA-seq data processed by *LIONS* following the standard protocol to generate raw calculation files (.pc.lcsv) which contain the input parameters used for classification (Figure 2.3). The starting simulation TSSs and *East Lions* analysis files were then parsed and merged in *R* and 2/3 of the cases (N = 77,775) were designated 'training set' and the remaining 1/3 of cases (N = 38,899) were hidden from the ANN model and designated 'testing set'. In such a strategy the ANN models are blind to the test data and a fair assessment of the models performance can be measured, this prevents 'over fitting' the classification model on the training data but failing to classify non-training data accurately.

The R package, *neuralnet [243]* was used to generate random starting neural networks using the resilient back-propagation with weight backtracking algorithm (rprop+) [244,245] for optimization of classification network based on the linear combination of the same parameters as the rational human algorithm. Classification for each of the three intersection cases (Up, UpEdge, and Einside) required a separate ANN model as the parameter profiles were distinct for these cases. ANN parameters were; random starting weights; 7 nodes in the input layer, 7 nodes in the hidden layer, one bias node, and one output node; cross-entropy error factor; 1e6 iterations per model; 0.0001 convergence threshold. Each model ran for ~200 CPU hours for a total of ~18,000 CPU hours of

directed training to yield the final three output *transcriptomeANN* models, selected as the best performing models on the test data.

2.2.7 Implementation

The core *LIONS* pipeline was written in *bash* script language. BAM analysis software was written in *Python3* (3.5.2). Data analysis and statistical calculations were performed by R statistical language (3.5.1). The source code for all *LIONS* components is available at www.github.com/ababaian/LIONS and all analyses are based on a standard .lions output file. A Docker container with *LIONS* installed is also available for virtualization.

File format standardization was performed to encourage users to share down-stream analysis scripts such that graphs and statistics of TEs could be reproducible and applied to different data sets quickly.

2.3 Results and discussion

2.3.1 LIONS

To quantify the contribution of TE promoters to the transcriptome from RNA-seq data alone, I was motivated to develop the *LIONS* analysis suite. Briefly, RNA-seq data along with a reference genome, gene and repeat annotation are inputs for the classification and annotation of TE-initiated transcripts (Figure 2.1A). For each RNA-seq library, a standard (.lion) file of TE-initiated transcripts is the output that can be grouped into biological categories such as cancer versus normal controls, for comparison (Figure 2.1B). A detailed outline of the analysis is provided in section 2.2.2. of the materials and methods.

TEs intersect exons in three main categories; as initiations at the 5’ end of a transcript; as exonizations either with or without being involved at a splice junction; and at the 3’ end as a termination site for transcripts (Figure 2.2A). The core *LIONS* classification segregates the

initiations from non-initiation events. This is biologically pertinent in the analysis of TE transcription since non-initiation events outnumber initiation events by three orders of magnitude (Figure 2.2B). Thus analyses based on chimeric read clusters alone, or TE-transcription levels alone do not necessarily reflect autonomous transcriptional activity of TEs but rather simply correlation or propensity to be transcribed as part of other transcripts. This is non-trivial as TEs have long been known to be enriched at 5' and 3' untranslated transcribed regions (UTRs) and within long-noncoding (lnc)RNAs [76,77].

2.3.2 Operating characteristics

To test the operating characteristics of *LIONS*, RNA-seq reads based on the ENCODE [246] K562 and H1 embryonic stem cell line transcriptomes were simulated at varying depths as a benchmark. Simulated TE-exon fragment clustering of reads plateaus at ~52% sensitivity regardless of further increase in sequencing depth (Figure 2.6A). This plateau emphasizes the systemic difficulty of accurately determining either 5' or 3' ends of transcripts from RNA-seq data alone, but the undetected TE start sites correlate with lower overall expression (Figure 2.6B). TE promoter analysis is confounded by the basic biological properties of TE TSSs, in that they are weaker and more biologically irreproducible (have higher cell-cell variation) than their non-TE TSS counterparts in CAGE analyses (Figure 2.7). From the fraction of TE TSSs which are measurable by chimeric fragments, the default *LIONS* parameters have a sensitivity of 36.35% and specificity of 98.63% (Figure 2.6C). The relative proportion of each class of TE TSS called largely matches the proportions of TE TSSs of the input transcriptomes, which rules out a systematic bias towards any one class of TE (Figure 2.6D). Altogether, while the set of TEs read-out by *LIONS* is not highly sensitive especially for lower expressed transcripts, it is highly specific and accurately reflects the underlying promoter activity of TEs.

In the context of cancer specific transcription these operating characteristics are quite favorable. It is a reasonable underlying assumption that genes which are biologically involved in oncogenesis will have relatively higher expression then non-functional or ‘noisy’ transcription, such as characteristic of TEs [75]. Since ~3% of all TE-exon interactions are TE-initiations, high specificity of the classification algorithm is important as for every one true positive TE-initiation case, there are 32 potential false-positives. The unequal distribution of positive and negative classification cases favors specificity for producing a reliable set of TE-initiations.

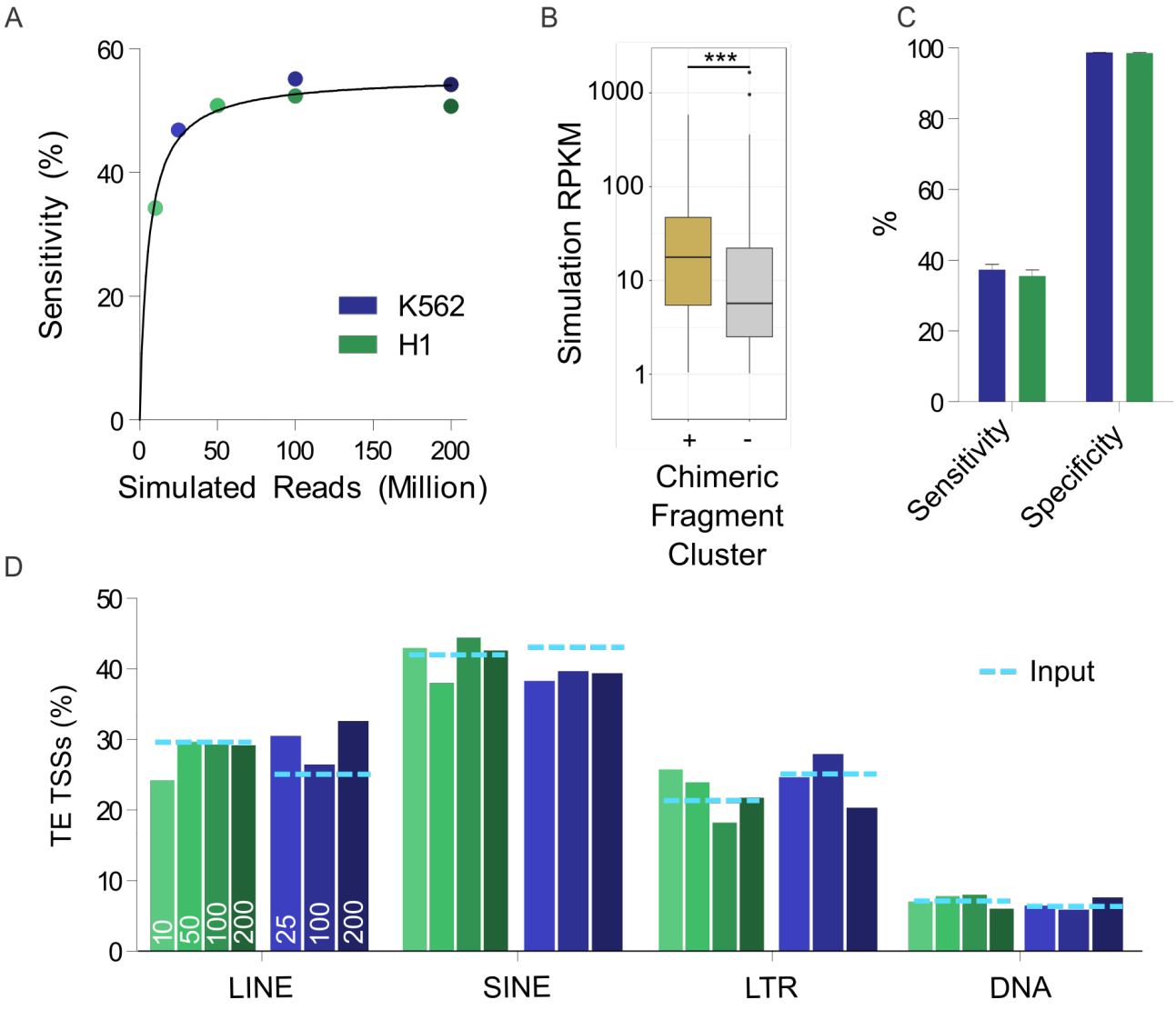


Figure 2.6: LIONS operating characteristics on simulated data

Simulated RNA-seq data based on a reference H1 ESC (green) and K562 (blue) transcriptomes were used as a benchmark to test the sensitivity and specificity of the LIONS suite. ***A)*** *In RNA-seq libraries simulated to varying depth, chimeric fragment clusters are limited in their capacity to detect TE-derived transcript start sites (TSSs), plateauing at ~52% sensitivity.* ***B)*** *The TE-TSSs that are detectable by chimeric fragment clusters (+) are more highly expressed (Welch's T-test, p = 4.59e-8) than those that lack chimeric fragment clusters (-).* ***C)*** *From the chimeric fragment cluster detectable TE-TSSs, default parameter LIONS has a 36.36% sensitivity and 98.63% specificity yielding a specific set of TE-TSSs.* ***D)*** *The relative proportion of LIONS called TE-initiated transcripts from each TE-class for each simulated data-sets at varying simulation depths, relative to their respective input transcriptome TE-class proportions (teal line).*

It should be noted that *LIONS* is dependent on an accurate reference genome, polymorphic or novel TE insertions are not reliably detectable. This is of most importance when considering the so called “hot” L1 transpositional activity in cancer, as any newly inserted L1 elements could initiate transcription from their bi-directional promoter. Overall the detection of LINEs is equivalent to non-LINEs (Figure 2.6), and since reverse transcription to the complete 5’ end of LINEs is rare, the promoter capacity of this class of LINEs is not expected to be a major source of error, but it may still be biologically significant to a patient.

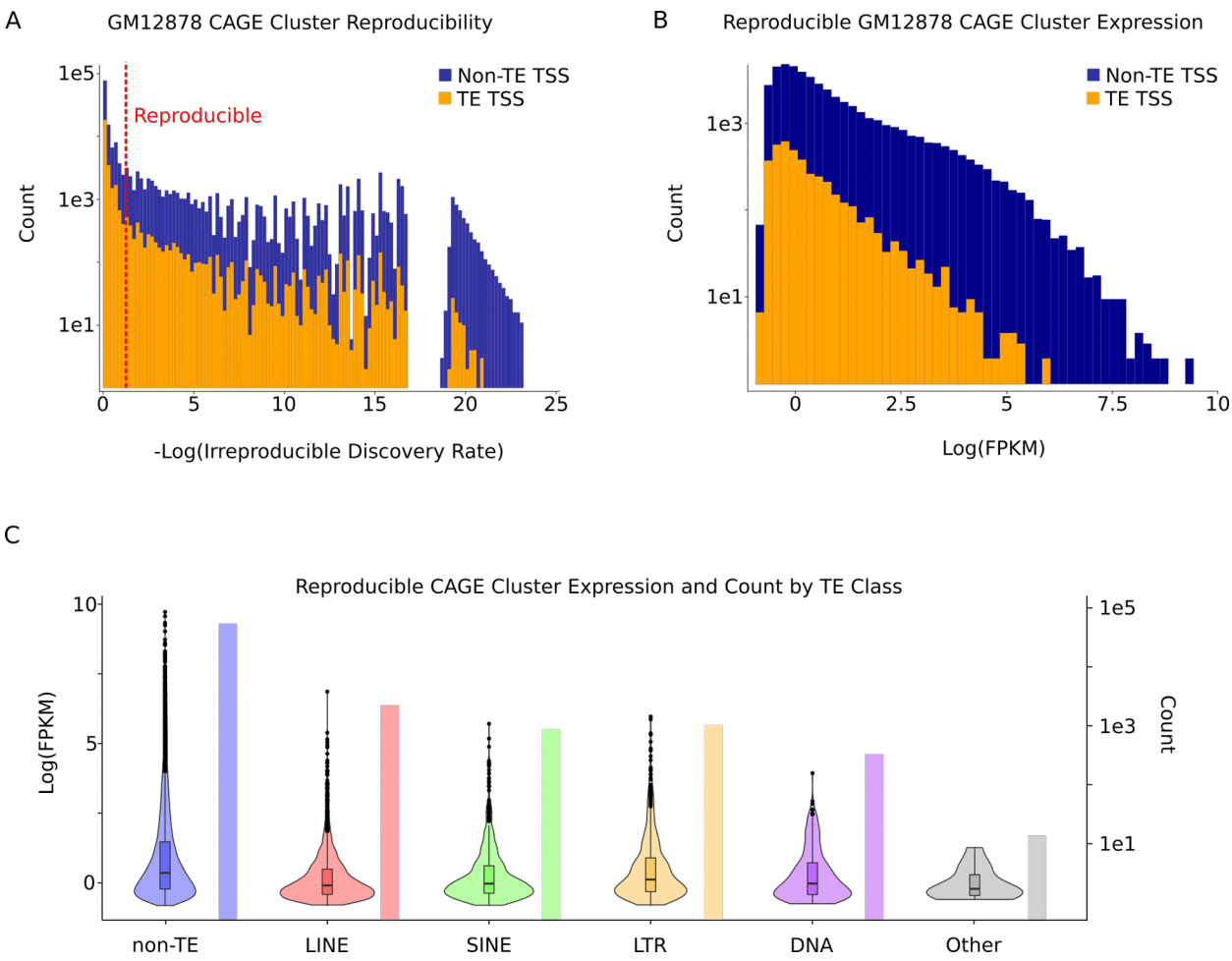


Figure 2.7: Reproducibility of transposable element (TE) transcription start sites by CAGE

5` Cap analysis gene expression (CAGE) transcription start site clusters were downloaded from the UCSC genome browser for GM12878 polyadenylated whole cell RNA (UCSC accession: wgEncodeEH001680). The center of each transcription start site (TSS) cluster was intersected against RepeatMasker to distinguish non-TE TSSs ((blue) and TE-TSSs (orange). A) To test if TE-derived TSSs are more or less variable between biological replicates the irreproducible discovery rate (IDR) between the groups was compared. TE-derived TSSs are more variable between biological replicates (Welch's t-test, p < 2.2e-16) then non-TE TSSs. Reproducible clusters are those that pass an IDR cut-off of <0.05 (right of red line). B) Among the reproducible CAGE clusters, TE-derived TSSs have a lower (Welch's t-test, p < 2.2e-16) expression level by log fragments per kilo-base per million mapped reads (FPKM). C) The TE-TSS clusters can further be striated by TE-class. Violin plot of the kernel density of the log(FPKM) is shown for each class overlaid with a bar graph of the count per TE-TSS.

To evaluate the accuracy of *LIONS*-classified TE-initiations on biological data and measure *in silico* and *in vivo* concordance, a set of Hodgkin lymphoma cell line specific and recurrent (relative to B-cell controls) RNA-seq data were analyzed by *LIONS*. Chimeric transcripts identified by *LIONS* were then assayed by RT-PCR on nucleic acids extracted from the respective cell lines. *In silico* predictions were largely in agreement with RNA assayed by RT-PCR at 70.7% and 89.5% sensitivity and specificity respectively (Figure 2.8).

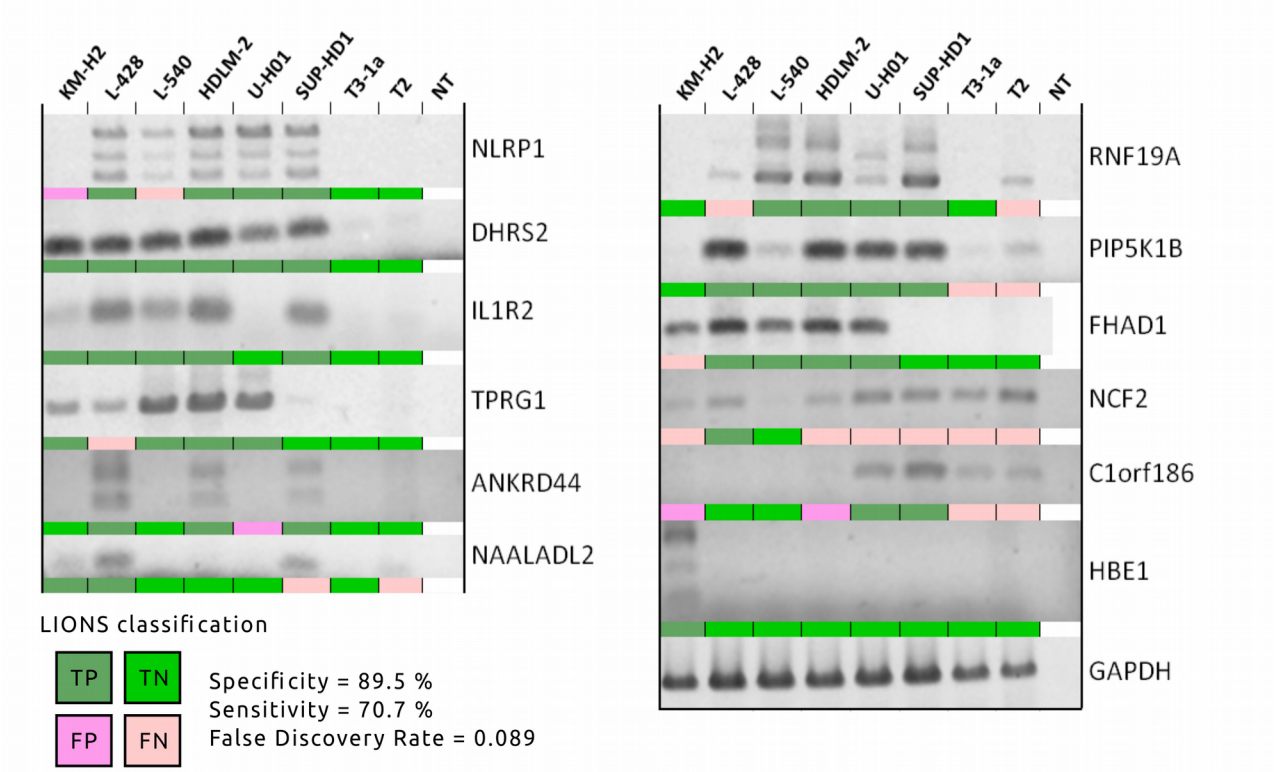


Figure 2.8: Reverse-transcription PCR validation of candidate TE-initiated transcripts

*From the Hodgkin Lymphoma (HL) RNA-seq data-sets, TE-initiated transcripts with predicted i*ntact coding sequences that occurred in at least 2/12 HL libraries and were absent from all nine primary B-cell libraries were selected as Hodgkin-specific and recurrent. Candidate genes were selected for potential involvement in cancer pathogenesis by a literature review. The TE-initiated isoforms were validated by reverse-transcription (RT-)PCR and compared to the in silico prediction from *LIONS*. The normal B-cell lines T2 and T3-1a were used as controls to test for HL specificity. A dark green bar indicates concordant detection between *LIONS* and RT-PCR (true positive), while light green indicates concordant absence (true negative). Magenta bars indicate *LIONS*-predicted and RT-PCR negative (false positive) and pink is the converse (false negative). RT-PCR is expectedly more sensitive for low-abundance transcripts (note the fainter bands in the false negative cases).

2.3.3 Artificial neural network classification

An algorithm based on human intervention to set the parameters is ultimately biased, so an alternative and arguably more empirical approach is to use machine learning to train *LIONS* on a simulated data set in which ground-truth is known. With this motivation the sub-package *Brunswick*, which trains and incorporates an artificial neural network (ANN) classifier, was developed and the *transcriptomeANN* mode for *LIONS* was implemented.

The RNA-seq simulation data was used for ANN training as this data was sufficiently abundant and is a defined 'ground truth' with respect to knowing what is a true TE-initiation event and what is a TE-exonization or TE-termination. *LIONS* was run on the simulated RNA-seq data and the raw calculation files (.pc.lcsv format) were used as input for ANN training and evaluation. Each TE-exon interaction containing a chimeric fragment in the simulated H1esc and K562 transcriptomes (N = 64,417 and N = 52,264 cases, respectively) was used for training and evaluation. Two thirds of the cases were randomly assigned as "training data" and one third was kept blind from the model and kept as "test data". The objective of the ANN was to distinguish the TE-initiation events (true positives) from TE-exonization and TE-termination events (false positives). Following a sum total of ~18,000 CPU training hours, 130/900 ANN models had converged on solutions.

An ANN architecture of seven input-layer nodes, seven hidden-layer nodes with a bias node which combine linearly into an output classification "TruePos" (Figure 2.9A) was chosen. It was inferred from manual classifications that the numerical requirements for the different TE-exon intersection cases (Up, UpEdge, and Einside) were different from one another, so to account for this, separate models were trained for each of the intersection cases.

The optimal ANN models showed strong classification receiver operating characteristics with the area under the curve (AUC) being 0.947, 0.868, and 0.837 for Up, UpEdge, and Einside, respectively (Figure 2.9). In the test data, the majority of cases fell under the Up intersection

category (450/749 cases) which also was the best performing model of the three. Overall weighting the models by the abundance of the cases, the ANN classifier had a sensitivity of 86.51% and specificity of 84.78%, which is markedly more sensitive but less specific than the manual classification of the same data-set at values of 36.36% and 98.63%, respectively.

These results are encouraging and offer a proof-of-concept that machine-learning approaches can be utilized for the classification of TE-initiation events. One immediate extension of this TE-initiation classifier would be to train complementary, TE-exonization and TE-termination classifiers. In this way each TE-exon interaction case is independently scored and analysis can be expanded to consider how cryptic sites in TEs influence transcript structure. In addition, these results could offer a generalizable strategy for *ab initio* sequence assembly, one in which specialized machine learning classifiers score the fidelity of individual components of a transcript assembly, such as transcription start site, splice junctions, or termination site, and these scores are used to refine contig assembly.

While the *Brunswick* classifier component performed well on the simulated RNA-seq data, when applied to biological RNA-seq data, the classifications were prone to errors in area of complex transcription. This is most likely due to simplicity of the simulated RNA-seq data, where the input transcripts are taken as ground truth and factors such as intron-retention, and transcriptional background are not modeled. As such, until a more biologically precise ground truth data-set could be defined, the output of any machine-learning based algorithms must be interpreted carefully.

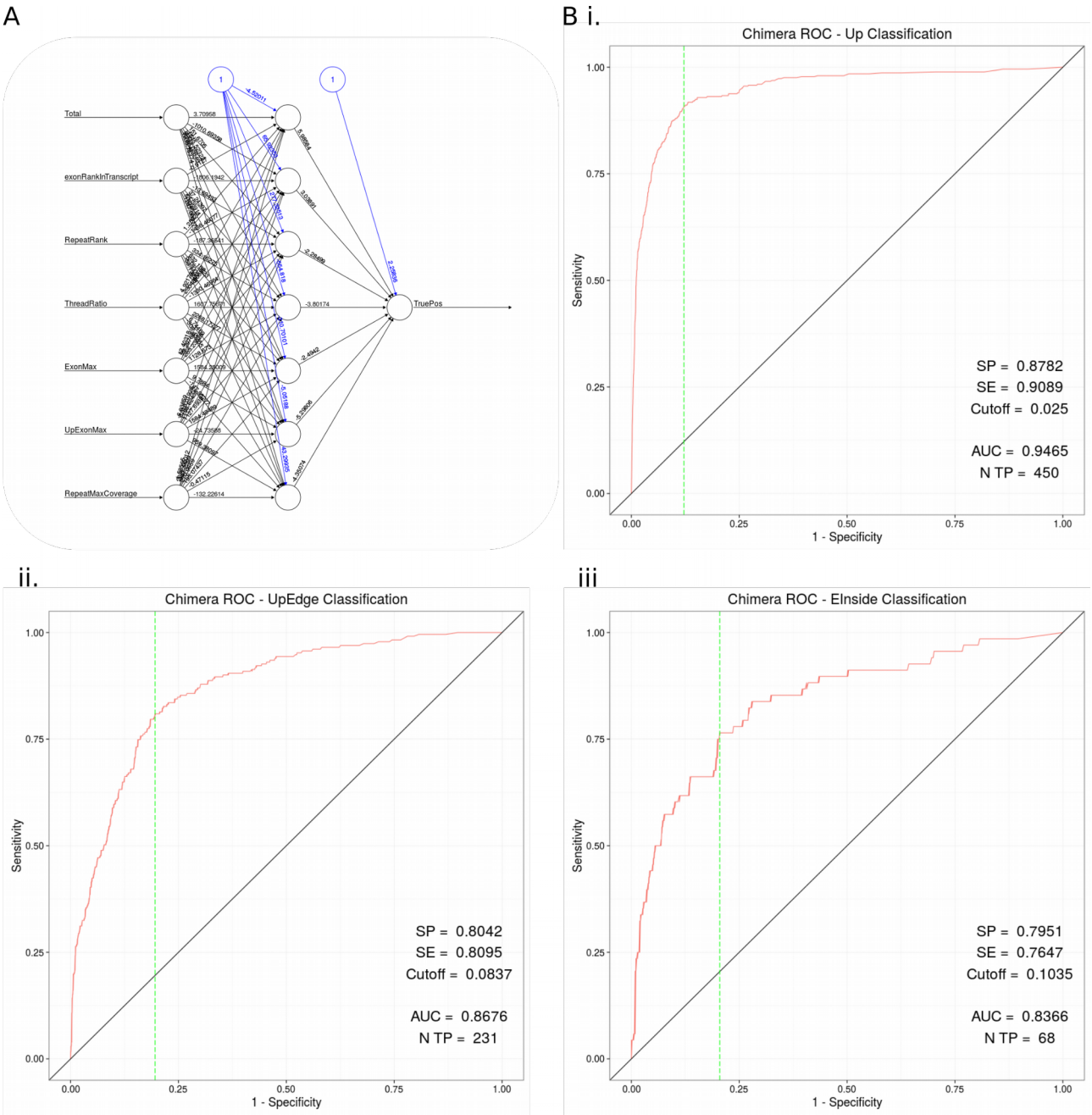


Figure 2.9: LIONS artificial neural network classifier

The transcriptomeANN mode of *LIONS* classifies TE-exon interactions as initiation or non-initiation using an artificial neural network classifier. **A)** Representative architecture of the ANN models showing the input, hidden and output layers. **B)** The receiver operating characteristic (ROC) curves for the three optimal ANN models for i. Up classifier, ii. UpEdge classifier, and iii. Einside classifier. The specificity (SP), sensitivity (SE) is reported at the output parameter cut-off selected as the minimal euclidean distance to the ideal (0, 1). The area under the curve (AUC) is reported as well as the number of true positive cases (N TP) upon which the evaluation was performed.

2.3.4 Future developments and conclusions

The preceding principles of local RNA sequencing analysis to distinguish TE-derived transcription initiation from exonization or termination can also be seen as a specific-case of the *ab initio* RNA-seq assembly problem. Local calculations used in *LIONS*, namely read threading and upstream coverage could be generalized to the entire transcriptome. Further refinement of these methods such as inclusion of aligned-strand bias measures [247], position-aware Hidden Markov Model or additional machine-learning trained sorting algorithms to detect the molecular signature of TSSs could be used to improve the accuracy of transcript assembly.

LIONS suite is limited in similar ways as other assembly methods are, namely in regions of high transcriptional complexity, especially if non-stranded data is used and there is bi-directional transcription. The coverage around all transcript ends in RNA-seq is reduced relative to interior sequences [247] and confounded by lower overall expression and higher variability of expression of TE TSSs in general [75].

One important consideration is that single-exon assembled contigs that initiate at a TE are explicitly excluded from further analysis by the sorting algorithm. This was an experimental design choice suited towards the application of *LIONS* for a higher specificity in detecting chimeric transcripts (TE-protein coding gene fusions) in cancer transcriptomes. Considering that LINEs and SINEs produce single-exon transcripts for native retrotransposition, this method will underestimate the transcriptional capacity of these elements, a measurement which is instead better performed by alignment to a consensus repeat sequence instead.

The focus of the *LIONS* suite on transcriptional initiation is the low-hanging fruit for TE-gene interactions. Additional analysis of chimeric read clusters may quickly yield TE sets which are incorporated into transcripts, such as TE-derived splice acceptors and donors in the newly classified characterized exitrons (also called retained introns, a sequence which can be both an exon and

intron)[248]. Anecdotally, one of the largest difficulties in developing *LIONS* was distinguishing the true initiation events from exitron-like events that occur within a TE. This distinction is also one of the greatest limitations of previous studies looking at TE-derived transcriptomes [230,232,234], which did not make this distinction.

Altogether *LIONS* is able to detect a specific set of TE-initiated transcripts from RNA-seq data alone. The detected set is enriched for higher expressed transcripts which, in a biological context such as cancer, are expected to be more relevant than the low expression / high variation TE-initiated transcripts.

Chapter 3: Transposable element promoters in cancer transcriptomes

3.1 Introduction

Although the concept of TE exaptation as a driving force in organismal evolution is becoming increasingly accepted [249] there is also interest in determining the potential role of TEs in human diseases, particularly cancer. Much of the recent focus has been on detection of new somatic insertions of L1 long interspersed elements (LINEs) in human malignancies [46,121] and on potential carcinogenic roles for HERV encoded proteins [142–144]. Newly integrated retroviruses have long been known to activate proto-oncogenes via the enhancers or promoters in their LTRs and, indeed, many of the most well studied oncogenes were originally discovered as common sites of retroviral insertion in animal cancer models [250]. It is possible that a similar process involving transcriptional activation of normally dormant TEs/LTRs in cancer cells could drive ectopic gene expression or transcription and contribute to somatic evolution of the malignant state – a phenomenon I've termed, "onco-exaptation" (Chapter 1). The plausibility of such a scenario is increased in cancer which can be associated with genome-wide DNA hypomethylation and epigenetic pertubation [251,252], and possibly an increased transcription of TEs which occurs as a result of this, relative to normal somatic cells [72,244,245].

In this chapter, I explore the occurrence and distribution of TE-initiated transcripts in a cell-senescence model system and a cohort of colorectal carcinoma (CRC) and patient-matched normal RNA-seq, with the objective of understanding the etiology underlying TE transcriptional activation.

3.2 Materials and methods

3.2.1 Data-sets

The triplicate MDAH041 primary-cell and replicative senescence, and transformed-cell and induced senescence RNA-seq data [255] was downloaded from the NCBI Gene Expression Omnibus (accession: GSE60340). The CRC and adjacent patient-matched normal epithelium RNA-seq data [256] was downloaded from European Genome-phenome archive (accession: EGAD00001000215).

3.2.2 *LIONS* and data analyses

To comprehensively examine TE promoter activation in cell senescence models and CRC, I applied the *LIONS* (Chapter 2) pipeline to the paired-end RNA-seq data to detect and quantify the TE-initiated transcripts in each library. Briefly, each RNA-seq library was aligned to human reference genome *hg19*, with *tophat2* (v.2.0.14) [237] and transcriptomes were assembled *ab initio* with *Cufflinks2* (v.2.2.1) [238]. The assembled contigs were then analyzed for evidence of overlap with RepeatMasker [12] annotated transposable elements to define TE-initiated transcripts (see: Chapter 2).

LIONS was run uniformly across all data-sets with the default ‘oncoexaptation’ parameters: *`crtReads='3'; crtThread='10'; crtDownThread='10'; crtRPKM='1'; crtContribution='0.1'; crtUpstreamCover='2’; crtUpstreamExonCover='1.5'`* (Figure 2.3).

Analysis of *LIONS* output data was performed with custom R scripts. Error bars shown on boxplots are 1.5 the inter quartile range, and on bar graphs the standard error of the mean, unless otherwise stated. Two-tailed Welch's t-test was performed to test for difference in the means with unequal variance using GraphPad Prism 5.0.3 for Windows (GraphPad software, La Jolla California USDA).

3.2.3 TE-initiation data simulations

For the empirical comparison of TE-initiation distribution to a random expectance, random distributions of TEs were generated. This was chosen to match as closely as possible, heterogeneity in the number of TE-initiations per library across the entire data-set.

For the *random spatial distribution* simulation, a random set of TEs were sampled without replacement for each simulated RNA-seq library, such that the total number of TEs sampled was equal to its respective CRC RNA-seq library. This process was repeated 1000 times independently to generate the empirical distribution.

For the *random recurrence distribution simulation*, all TE sites which were identified as active by *LIONS* in at least one library of the data-group was used as the input TE sample space. This excludes TEs which show no initiation activity in any data-set (zero occurrence). Each input TE was assigned randomly to one library to be present in at least one library. For each simulated library, a random TE set was then sampled from all input TEs without replacement such that the total number of TEs in the library matches its respective observed library. This process was repeated 1000 times independently to generate the empirical distribution.

3.3 Results and discussion

Having established and optimized a computational tool to detect TE-initiated transcripts from RNA-seq data (Chapter 2), I applied this method to a cell-senescence model system and a CRC and patient-matched adjacent normal biopsy data set.

3.3.1 TE promoter activation in senescent cells

A central tenant of this thesis is that an epigenomic dysregulation occurs in cancer that is necessary for transcriptional activity of TEs. Cancer versus normal comparison is between cells within a common cell lineage but from a separate individuals. In each cancer-normal pair, the cells

are separated replicatively (the number of cell divisions that have occured from the common stem cell of origin), through at least a single clonal expansion/bottleneck, and by major intrinsic cellular events such as cell crisis and/or transformation. To more finely understand how major transcriptional events such as, replication, senescence, and transformation can affect the transcriptional activity of TEs, I first investigated a model system in which several of these variables could be isolated.

The MDAH041 fibroblast cell line was isolated from a 22 year old female with Li-Fraumeni Syndrome (OMIM: #51623), an autosomal dominant pathology which predisposes patients to developing cancers at multiple sites including sarcomas, osteosarcomas, breast, brain and leukemias. Li-Fraumeni Syndrome is caused by inheritance of a heterozygous mutation in *TP53 (TP53$^{+/-}$)*, the most frequently mutated tumour suppressor gene across cancers [217,257]. The MDAH041 "normal" fibroblasts undergo spontaneous mutation in culture giving rise to an immortalized (*TP53$^{-/-}$)* cell line [258]. In the absence of mutation, MDAH041 cells will go through a set number of replications, and as this shortens the telomeres past a critical point (into non-telomeric sequence), the cells egress from the cell cycle into a state of senescence [255]. Alternatively, transformed MDAH041 cells can be forced into a state of senescence when treated with DNA damaging agents such as H_2O_2, 5-aza-2-deoxycytidine (5-aza), or adriamycin (generic name doxorubicin, a DNA intercalating agent) [255].

LIONS analysis was performed on a publicly available RNA-seq data set of replicative and induced senescence in MDAH041 cells, in triplicate for each condition [255]. Wildtype cells showed no difference in the number of TE-initiated transcripts between stable replication from passage 11 to passage 18, and by approximately passage 21, MDAH041 cells entered senescence (measured by beta-galactosidase activity in the original publication), and these cells have a marked

increase in TE-initiation (Figure 3.1A). The increase is driven by an increase in LTR transcriptional activity (Figure 3.1B).

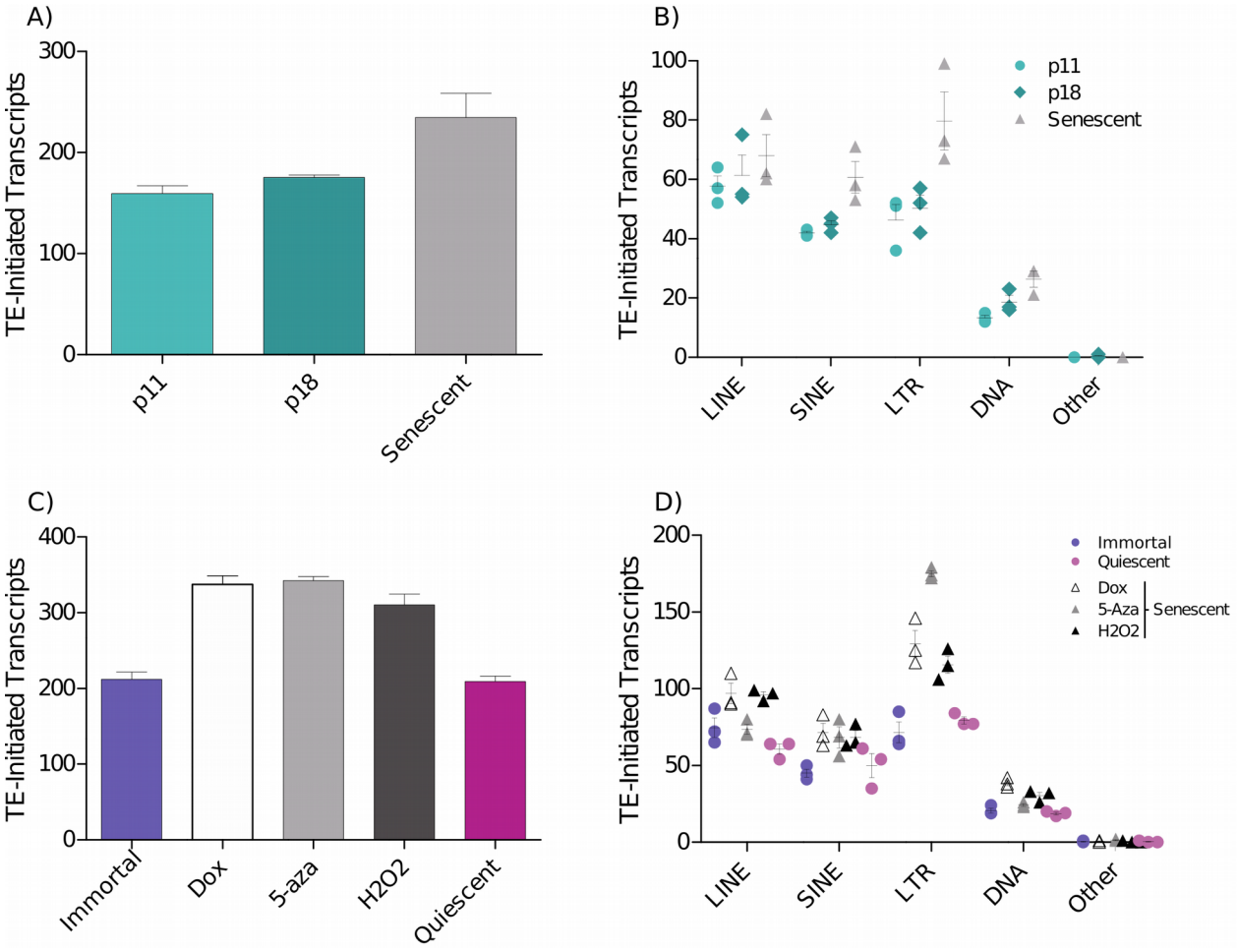


Figure 3.1: TE transcription in senescence

LIONS-classified TE-initiations in **A,B)** wildtype MDAH041 fibroblasts undergoing *in vitro* replication induced senescence and **C,D)** immortalized MDAH041 fibroblasts undergoing inductive senescence after treatment with doxorubicin (Dox), 5-azacytadine (5-aza) or peroxide (H_2O_2). As a non-replicative control, cells were serum-starved to induce quiescence.

Comparing wildtype and immortalized MDAH041 cells, the transformed cells have a greater level of TE-initiated transcription (Figure 3.1). Further, in the induced model of senescence from immortalized MDAH041 cells, all three senescence-inducing agents doxorubicin, 5-aza, and peroxide induced additional LTR transcriptional initiations. Most notably, 5-aza which results in

DNA demethylation had higher levels of LTR-transcription, even compared to doxorubicin and peroxide treatment. When the immortalized cells were serum-starved to force them into a non-replicative state of quiescence, this increase in TE-initiation was not seen suggesting this is not caused by exit from cell cycle but associated with the state of senescence specifically (Figure 3.1C).

Hierarchical clustering of the individual LTR loci active across the induced senescence data set recapitulated the treatment groupings (Figure 3.2A). The segregation of peroxide, doxorubicin and 5-aza induced senescence from one another in particular supports the idea that, while senescence leads to TE-initiation, the specific sub-set of TEs which become activated in each condition are more finely responsive to cell state. Perhaps unsurprisingly, 5-aza was the most responsive condition as DNA methylation is known to repress TEs, and the specific loss of DNA methylation by 5-aza activated a distinct set of TEs (Figure 3.2 B,C) [259].

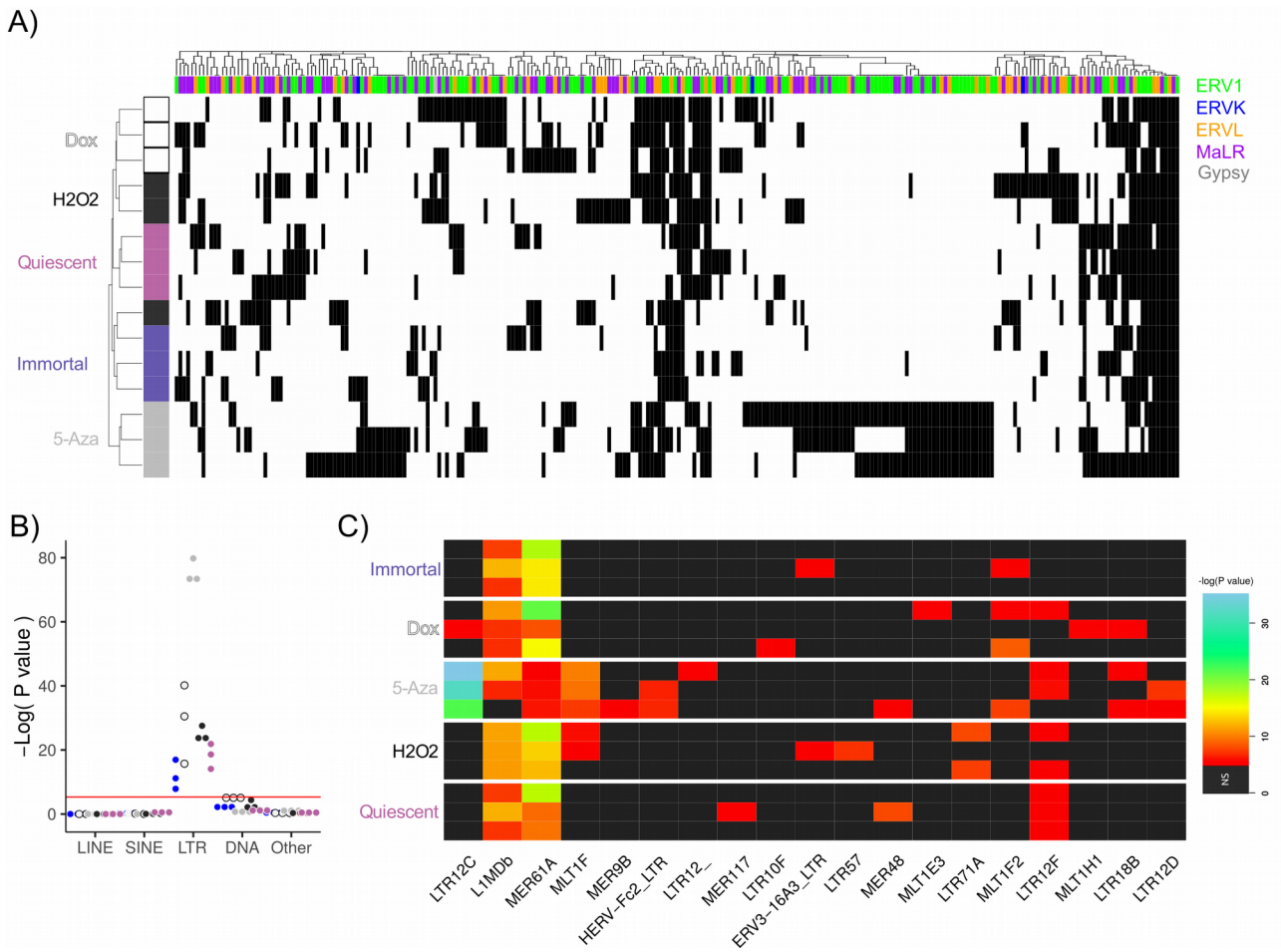


Figure 3.2: Clustering and representation of LTRs in induced senescence

A) Hierarchical clustering of all informative (present in >1 library) LTR-initiated transcripts identified in the induced senescence RNA-seq data. B) An exact binomial test of the relative over-abundance of each TE-class, normalized by all TE-initiations, the -log(p-value) for of each class is plotted. C) Similarly, a heatmap of the exact binomial test of the relative abundance of each particular TE-family, normalized by its respective TE-class.

To test if there was a particular family of TE which is enriched in senescence, a global TE Exact binomial test was performed for each TE class and family. Only the ERV1 class of LTRs was significantly enriched in 5-aza treatment (Exact Binomical Test, p = 0.0001). Across all conditions, MER61 LTR elements and L1MDb showed relatively high activity, implying these elements have a higher intrinsic activity in MDAH041 immortalized cells (Figure 3.2C). There was no specific TE family which showed reproducible enrichment across all senescence conditions. In 5-aza treatment

specifically, the LTR12C family, a relatively young and large LTR family, with notably high CpG density [260], was activated. Among the LTR12 family, the specific elements responsive to 5-aza were on average larger and more CpG dense then the genomic average for LTR12s (Figure 3.3). Altogether, TE transcriptional activation in senescence does not appear to be specific outside of LTR activation, with the exception of LTR12C activity in response to demethylation by 5-aza.

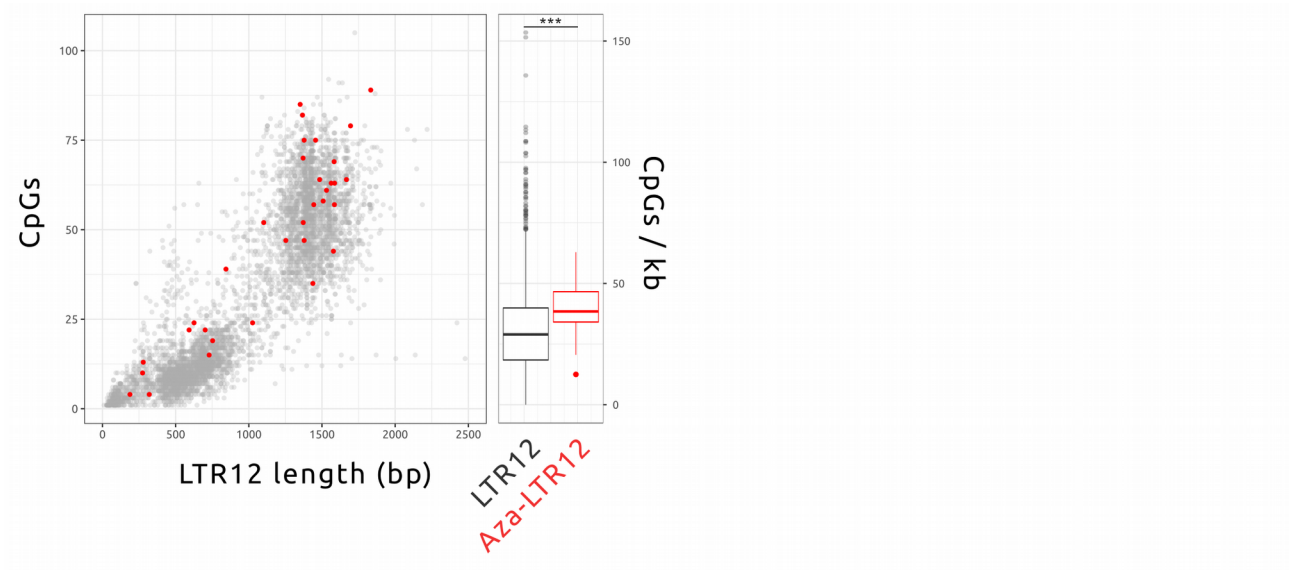


Figure 3.3: Length and CpG content of LTR12

For each LTR12 (LTR12, LTR12B, C, D, E, and F) locus annotated by RepeatMasker in hg19, the count of CpG dinucleotides and length of the element were extracted, and the density of CpG per kilobase were calculated. Individual LTR12s which initiated transcription upon 5-azacytidine (5-aza) treatment are shown in red. 5-aza responsive elements on average, contain more CpGs (48.0 vs. 32.1, Students T-test p < 1e-3), and are slightly longer (1177.5 vs. 1001.8 bp, p = 0.043) than the genomic average. Overall CpG density is also greater in 5-aza responsive LTR12s than genomic average (38.8 vs. 30.0 CpG per kilobase, p < 1e-4).

LTR12s (including LTR12B,C,D,E and F subtypes), which are the LTRs associated with the HERV-9 group [260], are much more numerous than other active ERVs, HERV-H or HERV-K, with solitary LTRs numbering over 7000 (Table 1.2). It is also a well studied HERV with several examples of LTR12s providing promoters for coding genes or lncRNAs in various normal tissues [99,261–264]. LTR12s, particularly LTR12C, are longer and more CpG rich than most other ERV

LTRs, possibly facilitating development of diverse inherent tissue-specificities and flexible combinations of TF binding sites, which may be less probable for other LTR types. Additionally, LTR12 elements are among the most enriched LTR types activated as promoters in HCC [147] and appear to be the most active LTR type in K562 cells [184].

LTR12-driven chimeric transcription in particular has been well documented [259]. One study specifically screened for and detected numerous LTR12-initiated transcripts in ENCODE cell lines, some of which extend over long genomic regions and emanate from bidirectional promoters within these LTRs [232]. The group of Dobbelstein discovered that a male germ line-specific form of the tumor suppressor *TP63* gene is driven by an LTR12C [263]. Interestingly, they found that this LTR is silenced in testicular cancer but reactivated upon treatment with histone deacetylase inhibitors (HDACi), which also induces apoptosis [263]. In follow-up studies, this group used 3' RACE to detect more genes controlled by LTR12s in primary human testis and in the GH testicular cancer cell line and reported hundreds of transcripts, including an isoform of *TNFRSF10B* which encodes the death receptor DR5 [149]. As with *TP63*, treating GH or other cancer cell lines with HDAC inhibitors such as trichostatin A activated expression of the LTR12-driven *TNFRSF10B* and some other LTR12-chimeric transcripts and induced apoptosis [149,265]. Therefore, in some cases, LTR-driven genes can have a proapoptotic role. In accord with this notion is a study reporting that LTR12 antisense U3 RNAs were expressed at higher levels in non-malignant versus malignant cells [266]. It was proposed that the antisense U3 RNA may act as a trap for the transcription factor NF-Y, known to bind LTR12s [267], and hence participate in cell cycle arrest [266].

The specific activity of LTR12 to 5-aza treatment in the cell senescence model, and numerous reports of activity in various cancers, raises the interesting possibility that this set of elements may be particularly responsive to the condition of genomic epigenetic derepression by DNA demethylation [268] or histone deacytlation [265]. It would be informative if counter-factual

evidence is found, testing on the genomic scale if demethylation is sufficient for an LTR12 response within a broad context of tissues, or if this occurs under additional molecular prerequisites met in fibroblasts and germ cells.

In a recent analysis of the same Purcell et al., data of induced and replicative senescence [255], Colombo et al., reported, in agreement with my analysis, that the overall transcriptomic contribution of TEs correlates universally with senescence induction [269]. Our data contrast in that they observe the largest transcriptional induction in the LINE L1HS, and L1PA3 family of elements, which highlights a difference between the methods. *LIONS* analyses consider the binary activity of individual TE loci, and the holistic initiation capacity of a TE group, whereas TE differential expression analysis can be strongly biased by a few hot loci, or be confounded by exonization which contributes several orders of magnitude more TE-derived reads (Figure 2.2B).

3.3.2 TE promoter distribution in crc and adjacent normal epithelium

There are two extremes which can model cancer-associated TE transcriptional activation. *1) The Stochastic Model*: TE activation is a random process across the genome, with each locus having a fixed and low probability of activation. In turn, measured increase in TE-initiated transcription reflects an underlying genome-scale phenomenon resulting in the dispersed activation of individual elements. *2) Deterministic Model:* the specific set of transcriptionally active TEs is a direct consequence of instantaneous cell-state. In turn, increases in TE-initiated transcription is caused by a specific change in cell conditions (such as transcription factor abundance), which leads to the programmed and deterministic activation of responsive elements. Most likely, both models have some truth in describing TE activation, but quantifying the relative contribution of each model has important consequences in understanding the etiology of cancer-associated TE transcriptional activity.

From a statistical perspective, this problem can be stated as, “What is the *clustering tendency* of TE transcription?” This can be interpreted both as the spatial genomic clustering, the distribution across linear chromosomes, and perhaps more meaningfully as the regulatory clustering, the distribution or correlation of TE activity across cell-states.

To best address this problem, a large cohort (n = 66) of RNA-seq libraries from Colorectal Carcinoma (CRC) biopsies and adjacent normal biopsy controls was used [256]. This data-set not only provides sufficient statistical power, it has a matched number of normal and cancer samples, all from the same patients which will account for patient-level variability.

Similar to cell-senescence, the CRC increase in TE-initiated transcription is the result of higher LTR-initiated transcription (Figure 3.4A). Comparing the patient-matched difference in TE-initiated transcripts between CRC and normal samples, the mean change of LINEs, SINEs, and DNA elements did not deviate from zero, while the CRCs gained on average 11.12 LTR initiated transcripts relative to their respective normal controls (Figure 3.4Aii). Unlike the cell senescence data where LTR12C/ERV1 was enriched in the 5’aza treatment group, no LTR class shows consistent statistical over-representation in CRC or normal RNA-seq (Figure 3.4B).

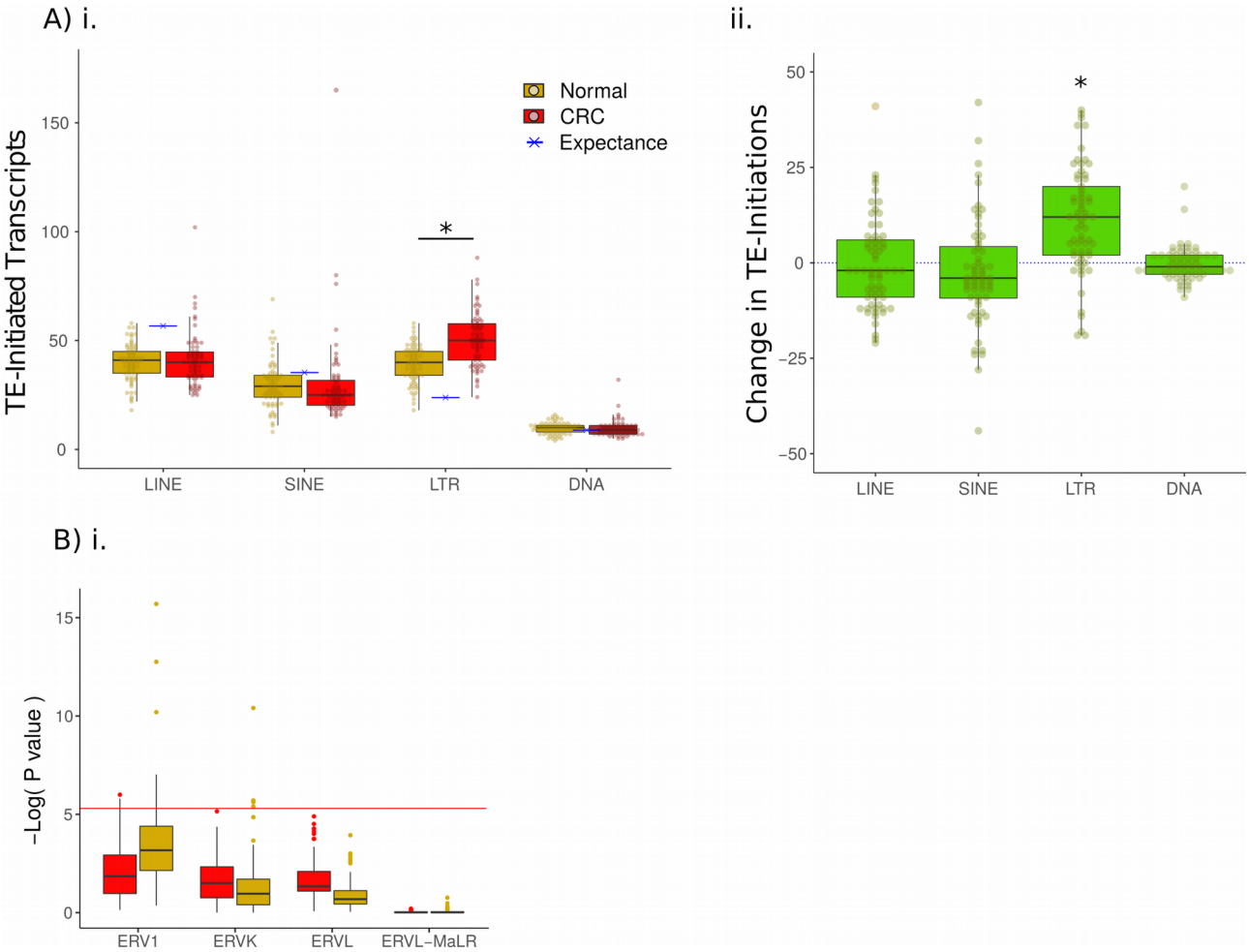


Figure 3.4: TE-initiated transcripts in CRC and adjacent normal

LIONS analysis of 66 patient-matched colorectal carcinoma (red, CRC) and adjacent matched normal epithelium biopsies (gold). A) i. The total number of initiations per TE family and ii. the difference (cancer – normal) in family TE-initiations between patient-matched CRC and normals. LTR elements are increased in CRC (p = 1.63e-8, two-tailed t-test). B) An exact binomial test of the relative over-abundance of each TE-class, normalized by all TE-initiations,the -log P-value for of each class is plotted. Red horizontal line demarcates a multiple-testing adjusted p = 0.05 significance level.

To test for spatial clustering of TE-initiations along chromosomes, for each library, the minimum distance in base-pairs between two TE-initiations was calculated. In addition, 1000 random TE-initiation data sets were generated such that each set contains the same number of samples with matching number of TE-initiations per sample as in the cancer set (Figure 3.5A). There was no difference in the mean distance between TE-initiations in pairwise comparisons

between CRC vs. Normal, CRC vs. Random or Normal vs. Random (Figure 3.5Aii), while CRC and Normal sets do have more TE-initiations relative to Random between 1-10 kb of one another, at 359, 251, and 79.3 (+- 11.3) initiations, respectively. This spatial clustering likely represents a modest increase in the probability of TE-transcriptional activation occurring within already open/transcribing chromatin domains. One assumption underlying this analysis is that the samples are approximately karyotypically normal, a reasonable assumption for the normal tissues, but almost certainly not true for CRC, especially microsatelite unstable samples [270]. As such the modest increase in TE spatial clustering in CRC relative to Normal is likely too conservative and a more accurate measurement would require matching genome/transcriptome assemblies.

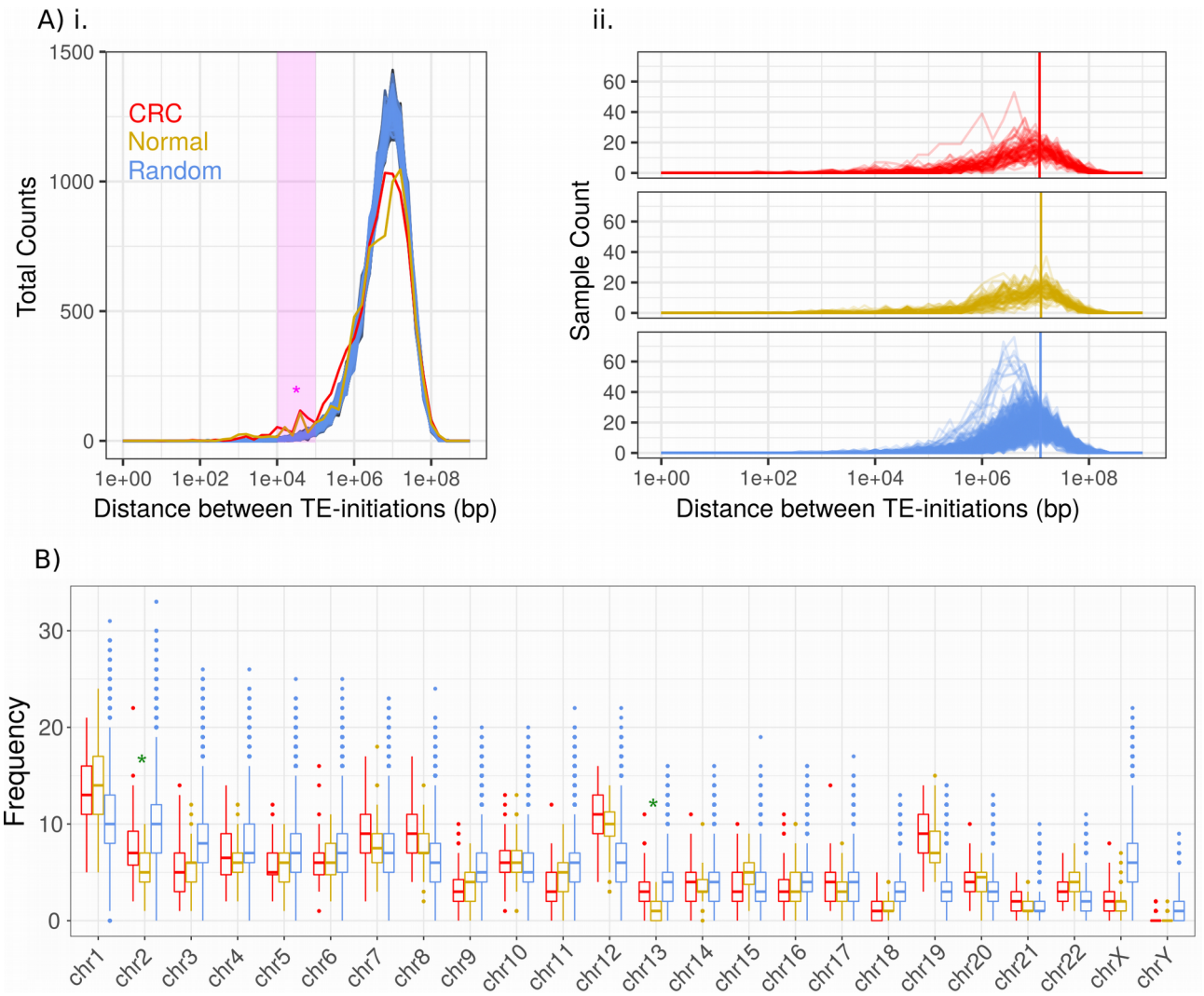


Figure 3.5: Spatial clustering of TE-initiations in colorectal carcinoma

*T*he spatial distribution of *LIONS* classified TE-initiations along chromosomes in colorectal carcinoma (CRC, red), patient-matched adjacent normal tissue (gold, N = 66) and 1000 sets with randomly distributed TEs (blue). **A) i.** For each TE-initiation, the minimal distance to the next closest TE-initiation was calculated. The frequency plot shows the total counts across all samples in its biological group. CRC and Normal samples both contain an enrichment of TE-initiations between 1e4 and 1e5 bases apart (purple highlight). **ii.** The same data deconvoluted to show the frequency per sample and mean distance (vertical line), only 10 Random sub-sets are plotted to prevent overplotting. **B)** The distribution of TE-initiations across each chromosome. There is a difference in mean frequency/ chromosome between the CRC and Normal on chromosome 2, 13, and 15 (p adj. = 1.47e-4, 5.59e-7, and 2.49e-2 respectively, yellow star, Welch's Two Sample t-test with Bonferroni correction).

The distribution of TE-initiations between chromosomes was within the range of 1000 random simulations, meaning, empirically no chromosome has more initiations than expected at an empirical $p < 0.001$, but the mean number of observed TE-initiations does deviate from random mean (Bonferonni-adjusted Welch's Two Sample t-test, $p < 0.05$) on all but chromosome 6, 10, 14, 16 and 21 (Figure 3.5B). Most notably is the increase in TE-initiations on chromosome 12 and 19, and depletion on the sex chromosomes. Comparing the mean number of TE-initiations per chromosome between CRC and Normal libraries, CRC contains significantly more TE-initiations on chromosome 2 and 13, and is depleted for initiations on chromosome 15. These differences arise from the deviation of the normal libraries relative to the random-expectancy which is reversed in CRC suggesting that in normal cells chromosome 2 and 13 harbor particularly repressed TEs and chromosome 15 permissive TEs, although the total number of events per chromosome remains at a moderate level.

Distinct from the spatial clustering of TEs across the genome, the regulatory or co-occurrence clustering of TEs can be considered. In this context of clustering tendency, LTR activation demonstrates that there is non-random TE-activation, certain elements namely LTRs have a higher activation probability relative to LINEs, SINES or DNA TEs (Figure 3.4A). As previously discussed, the intrinsic promoter capacity of LTRs is expected to be higher than other TE classes as LTRs evolved to function as promoters in ERVs. In addition, the mutation and regulatory degradation of elements is not expected to be equal across all TEs or all LTRs. Human LTRs range from 80 ka to >100 Ma in age, and as such vary in their state of decay. To account for this confounding variable, subsequent recurrence clustering analysis was limited to the set of TEs which initiate transcription in at least one sample.

To test if individual TE-initiation events are non-randomly distributed with respect to their occurrence frequencies from the sub-set of putatively active TEs, the recurrence of each TE-

initiation locus in CRC or normal controls was plotted (Figure 3.6A). The data was compared to a randomly simulated data-set with the same total set of TE-initiations and same number of TE-initiations per library as the data (Figure 3.6Aii). TE-initiation site distribution is strongly non-random when compared to simulated data. This suggests a high degree of heterogeneity in the activation potential of individual sites, with sites active in both CRC and normal (along the xy-axis), sites that are specific to normal samples (along the x-axis) and sites that are specific to CRC (along the y-axis). This demonstrates that at least a sub-set of TE-initiation responses are also condition specific. In contrast, when CRC data alone was sub-set and compared against itself, the tails along x and y-axis are absent (Figure 3.6B).

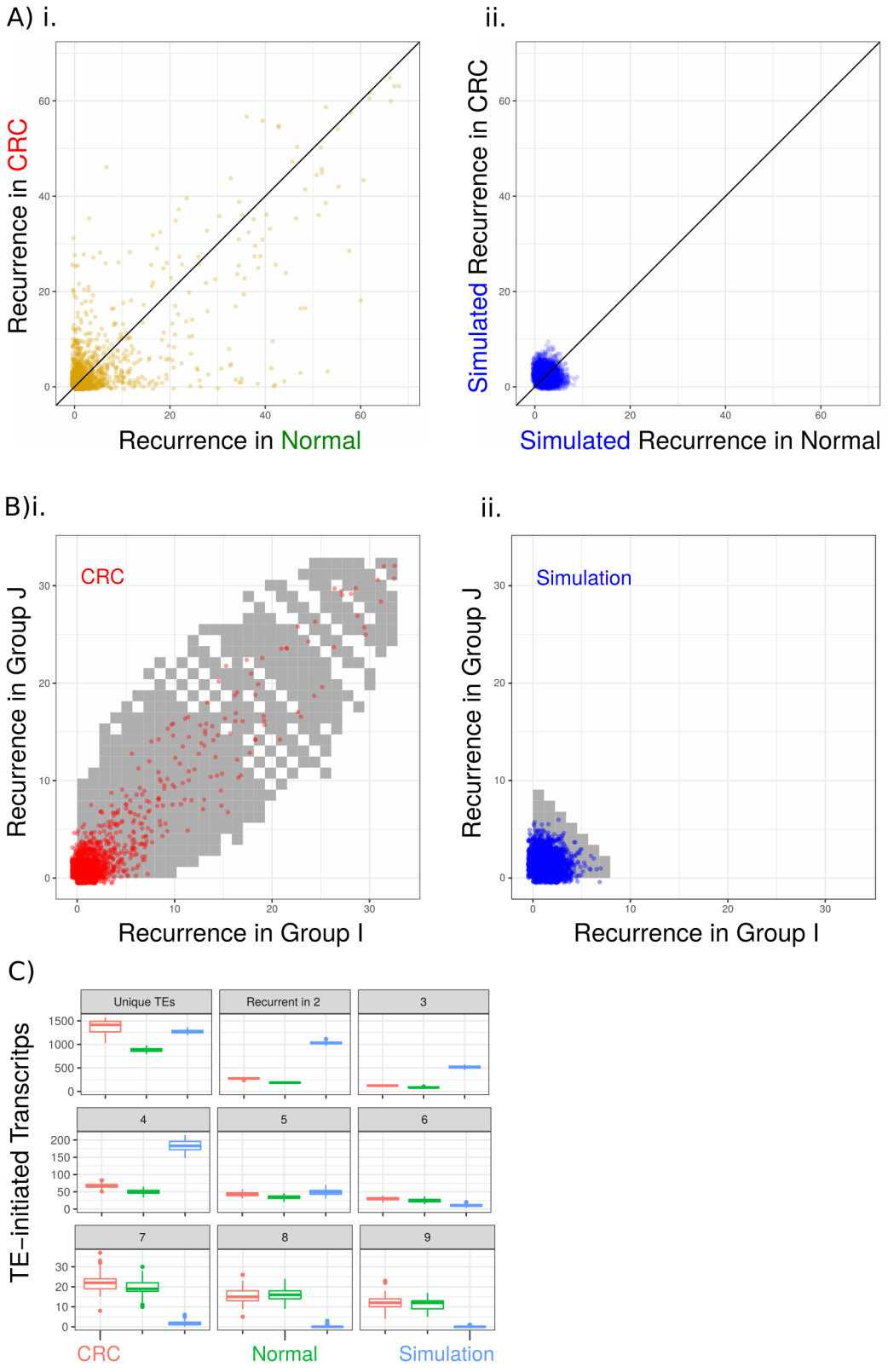


Figure 3.6: Recurrence of TE-initiations in CRC

Text 4: Figure 3.6 Continued.

Comparison of the recurrence of individual TE-loci in colorectal carcinoma (CRC, red), patient-matched normal controls (green) or a random simulation of TE-initiations (blue). ***A)*** *The inter-group recurrence of TE-initiation loci where* ***i.*** *each initiation locus present in at least one library is plotted as a point, showing how often each locus initiates transcription in each respective group* ***ii.*** *Simulated CRC and Normal data was generated for empirical comparison to observed data. Each simulated library randomly sampled an equivalent number of TE-initiations as its respective observed data from the the same total set of unique TE-initiating loci as in the observed data.* ***B) i.*** *The intra-group recurrence of TE-initiation loci where* ***i.*** *the data or* ***ii.*** *simulated data was randomly sub-divided into two groups for comparison (bootstrapping). The points show one bootstrap iteration and the gray shading shows the range from 100 bootstrap iterations.* ***C)*** *The total number of TE-initiations that are unique to one library, recurrent in exactly two, or three, …, or nine libraries. Distribution were generated by down-sampling of the data to 45 randomly selected libraries each.*

Counter intuitive to the high recurrence values, CRC TE-initiation loci that are unique to a single library from all CRC libraries, are more abundant than unique Normal TE-initiations or simulated TE-initiations (based on the CRC data-set). In this way, TE-initiations in CRC are both over-dispersed at the level of unique sites, and highly-recurrent at least 7 or more (>10%) libraries (Figure 3.6C).

What this means is that the majority of TE-activation space in CRC are unique activations distributed across many elements and this supports a model where TE-activation in CRC is highly noisy. The same level of unique activation is not seen in the normal controls when compared to the simulation. Simultaneously, a small sub-set of elements in both CRC and normal, are highly recurrently activated. Interpreting TE-activation as a cell-specific response, it is expected to see a sub-set of elements be highly recurrent to CRC or normal since these cells share a transcriptional program. The over-dispersion of unique elements in CRC is unexpected but provides a key insight into transcriptional innovation in cancer. Normal cells share a common differentiation path, reflected by gene expression patterns. Cancer cells share some common hallmarks during oncogenesis, but the path by which they reach their current state is unique. The unique

transcriptional histories of each cancer is reflected in the abundant unique TE activity. Thus (speculatively), the level of unique TE activity is proportional to the divergence of the transcriptional programs of two cells.

3.4 Insight into TE-initiated transcription

Altogether, these data on the distribution of TE transcriptional initiations in CRC and cell senescence provide insight into the underlying nature of this phenomenon. There are two broad types of TE-initiation loci, *unique sites* with rare activation potential, and these make up the majority of active loci; and *recurrent sites*, those that are informative of a particular tissue and/or cellular state. What this implies about the etiology of TE-initiated transcription is that there are likely two mechanisms by which they arise.

The over-dispersion of unique sites is consistent with the idea of ‘transcriptional noise’ for TE-initiated transcription. These numerous elements (recall, there are >800,000 LTR fragments alone in the human genome), have a low probability of activation, and during the course of an individual organism’s development or an individual cell lineage’s development, rare activation (with respect to the population) give the individual a unique ‘transcriptional fingerprint’ of TEs. This activity is likely to be heterogeneous at multiple levels of analysis: across single-cells, across tissue, across individuals and even possibly across genetically diverse populations, although additional research is needed to address each of these questions in turn. The consequence of such transcriptional diversity is fascinating to speculate about. TE-initiated transcription doesn’t have the obvious evolutionary constrictions as native-gene promoters, and as such could be a mechanism of generating phenotypic diversity, even among closely related individuals by varying the gene expression of neighboring genes, making it an intrinsic epigenetic mutagen, with the potential for generating both negative and beneficial variation.

The *recurrent sites*, like those over-represented at >10% samples as in CRC, have a higher per-sample activation rate by definition. This set of elements has been described in CAGE- and RNA-seq based analyses as being highly tissue-specific [67,75,221], or condition-specific such as 5-aza responsive elements in cell senescence. In this case, the set of tissue-recurrent TEs can be interpreted as a noisy reporter or even a classifier of instantaneous cell state. It would be intriguing to analyze much larger and diverse RNA-seq data sets, building up a repertoire of tissue and condition-responsive TEs. With such a data set, it would be possible to determine 'TE condition signatures'. These would be quite similar to empirically derived gene expression signatures, but without the constriction that each signal is a component in a larger transcriptional program, meaning each signal is a (more) unbiased reporter of the condition. This method would lack obvious functional information as a gene set contains, but would be an empirical correlative. Where this becomes further relevant are cases in which TE *recurrent sites* are not neutral bystanders to the transcriptome, but confer *de novo* function to cells, which in cancer are referred to as onco-exaptation cases, and are explored in the following chapter.

Altogether, this sketches a picture of TE-initiations as a highly stochastic and cell-specific process, with a sub-set of conditional response elements.

Chapter 4: Transposable elements mediated transcriptional innovation in lymphoma

4.1 Introduction

Classical Hodgkin lymphoma (cHL) is one of the most frequent lymphomas in North America [271–273] and, while prognosis is generally favorable, ~20% of patients still die of this disease. The malignant cells of classical HL, the Hodgkin Reed-Sternberg (HRS) cells, are derived from germinal or post-germinal center B cells [273]. Unlike other lymphomas, HRS cells have undergone major reprogramming of gene expression with loss of expression of most B-cell specific genes and gain of expression of genes normally active in other hematopoietic cells [274]. Both global epigenetic changes and deregulated transcription factors, such as NF-KB, are involved in reprogramming of gene expression and transformation to malignancy in HL (reviewed in [275]).

cHL is unique among cancers because the malignant HRS cells comprise only <1% of the tumor, making them difficult to interrogate experimentally or monitor in a patients at the molecular level. These rare HRS cells coordinate a permissive tumor microenvironment, promote malignant growth and immune evasion [276]. Despite the improvement in cHL treatment, patient outcomes [277], and understanding of its pathobiology [278], there remains unmet prognostic needs [279]. In particular, accurate prognostics and/or predictive biomarkers are needed to inform decision making at initial diagnosis to: (i) identify patients at risk of relapse and requiring upfront aggressive therapies such as hematopoietic stem-cell transplantation, and (ii) identify patients with favorable prognosis for milder treatments in this young cohort to mitigate the long-term harm caused by standard therapies, such as cardiotoxicity [280].

Diffuse Large B-cell Lymphoma (DLBCL) is an aggressive form of non-Hodgkin lymphoma, making up ~40% of lymphoma diagnoses [281]. DLBCL is broadly classified into germinal centre

B-cell (GCB) and activated peripheral blood B-cell (ABC) subtypes based on the similarity of the tumour to its developmental cell of origin [282] which is reflected in the global transcriptomic differences between these two subtypes [283]. In contrast to cHL, DLBCL cells are abundant within a tumour and primary patient biopsies can be readily analyzed by RNA-seq [179].

In this chapter, (i) I analyze the TE-initiation capacity of HL cell lines, diffuse large B-cell lymphoma primary patients and normal germinal B-cell biopsy controls. (ii) I characterize a HL-specific TE-initiated transcript, LOR1a-*IRF5*. (iii) I explore the applicability of using TE-initiated transcripts as a prognostic biomarker for cHL.

4.2 Materials and methods

4.2.1 RNA-seq alignment and analysis

The cHL cell lines and B-cell RNA-seq libraries (Supplementary Table 2.1) were hg19 aligned and analyzed by *LIONS* as described in Chapter 2 and 3. Sequencing coverage and genome browser snapshots were visualized on the UCSC Genome Browser [284].

Promoter contributions of LTR and native first exons to *IRF5* were calculated by defining all known *IRF5* exons from RefSeq, RACE and *in silico* assembly, and creating a custom reference map of all possible splice junction combinations. RNA-seq reads were then aligned using *bowtie2* [285] to the splice junction map. The coverage at the splice junction for each promoter-exon pair was summed to measure the relative LTR:Native promoter contribution to overall expression measured in reads per kilobase per million (RPKM). Code is available at *https://github.com/ababaian/Cypress*.

4.2.2 Cell culture

Cell lines, KM-H2 (cat#: ACC-8), L540 (ACC-72), U-HO1 (ACC-626) , L1236 (ACC-530), L428 (ACC-197) were received from the C. Steidl lab whom received them from DSMZ cell

repository (Leibniz, Germany) and previously validated the cell lines by karyotype and RNA-seq SNP analyses.

Cell lines were cultured under conditions recommended by DSMZ. Briefly, KM-H2, L1236 and L428 were cultured in 90% RPMI 1640 (RPMI, STEMCELL Technologies. Vancouver, BC. Cat#: 36750) + 10% fetal bovine serum (FBS, Gibco Laboratories. Gaithersburg, MD. cat# 12483-020). L540 were cultured in 80% RPMI + 20% FBS. U-H01 was cultured in 64% Iscove's MDM (IMDM, STEMCELL, cat#: 36150) + 20% FBS. All media was supplemented with 100 units penicillin-streptomycin (Gibco, cat#: 15140-122).

4.2.3 RNA and protein assays

For the preparation of protein lysates, $2x10^6$ cells were washed thrice in PBS, re-suspended into 100 µl RIPA buffer (Sodium Deoxycholate 0.5%, Ipegal, 0.01%, SDS 10% in PBS) with Complete protease inhibitor (Roche), homogenized by aspirating through at 21G syringe, incubated on ice for 10 minutes and immediately stored at -80°C. Upon thawing cell lysates, protein concentration was measured with Bradford reagent (Bio-Rad) and the colometric reaction was measured at 570 nm by Elx808 microplate reader (BioTek). For gel electrophoresis, equal protein were loaded in each lane, ran using the 4-12% Bis-Tris gels and the NuPAGE SDS-PAGE gel system (Life Technologies) and transferred onto PVDF membranes (Millipore). Membranes were blocked with 5% skim milk in TBST for an hour and cut to expected bands. IRF5 and Actin were detected with anti-IRF5 mouse monoclonal (1:1500, Abnova Taipei City, Taiwan. cat#: 2E3-1A11) and anti-Actin rabbit polyclonal (1:1000, Abcam Cambridge, UK. cat#: ab8227) antibodies, following overnight incubation at 4°C. Secondary antibodies incubations were goat anti-mouse-horse radish peroxidase (1:10000, Santa Cruz Biotechnology, Santa-Cruz, CA. cat#: sc-2005) and goat anti-rabbit-HRP (1:10000, Santa Cruz Biotechnology, sc-2030) for 1 hour at room temperature. Protein was visualized with

Amersham ECL Western Blotting Analysis System (GE) and developed on BioMax MR Film (Kodak). Protein band intensity quantification was performed with ImageJ software [286] and the ratio of IRF5 to Actin is shown below each lane. Blots were performed in duplicate.

DNA and RNA was simultaneously extracted from Hodgkin's cell lines (HDLM-2, KM-H2, L428, L540, L591, L1236, Med-B1 and UH-01) using the Allprep kit (Qiagen) and by the same method from non-Hodgkin's cell lines (GM12878, HL60, IM9, Jurkat, K562, NK92, Raji and THP1) provided by M. Romanish. Nucleic acids were quantified by spectroscopy with a Nanodrop 1000 spectrophotometer (Thermo Scientific). 1 ug of RNA was reverse transcribed using the VILO RT system (Invitrogen) unless otherwise stated.

Quantitative RT-PCR was performed on cDNA from different HL lines to assess the relative expression level of native 'a' isoform of *IRF5* and LTR-initiated LOR1a-*IRF5*. Total *IRF5* levels were measured using primers targeting exons 2 and 3 (the exons are downstream of both promoters converging). The relative promoter activity was measured as a ratio of LTR- to Native-specific transcription. Quantification performed using the delta-delta CT method [287] relative to *ACTB* levels. Primers are listed in (Supplementary Table 4.2).

To determine if the LTR element is truly the transcription initiation site of *IRF5* in HL, Rapid Amplification of cDNA Ends (5` RACE) was employed (FirstChoice RLM kit, Ambion Life Technologies, Grand Island, NY) kit with Superscript III (Invitrogen Life Technologies) polymerase used for reverse transcription and Sanger sequencing (Eurofins MWG, Ebersberg Germany). In L428, transcription initiated from within the LOR1a LTR and at both the native "a" and "d" start sites, as well as five other minor transcription start sites not previously characterized (Figure 4.3). The UHO1 cell line, which is negative for the LOR1a-*IRF5* LTR isoform by RNA-seq was used as a negative control.

4.2.4 DNA methylation analysis

Bisulphite sequencing was performed as previously described [288]. Briefly, 500ng genomic DNA using the EZ DNA Methylation Kit (Zymo Research) according to the manufacturer's protocol. Converted DNA was used as a template for 35 cycles of 1 round or 2 rounds in a semi-nested PCR reaction with AmpliTaq Gold DNA polymerase (Applied Biosystems). PCR was performed in duplicate. PCR products were gel-purified (Minelute) and cloned using the pGEM-T Easy Vector kit (Promega). All sequences included in the analyses either displayed unique methylation patterns or unique C to T non-conversion errors (remaining C’s not belonging to a CpG dinucleotide) after bisulfite treatment of the genomic DNA. This avoids considering several PCR amplified sequences resulting from the same template molecule. All CpH sequences had a conversion rate >96%. Plasmid preparation and DNA sequencing were performed by Eurofins MWG Operon. At least six independent clones were obtained for each region of interest. Data analysis was performed using the QUMA analysis program from RIKEN.

4.2.5 Microarray analysis and the HL-LTR NanoString assay

Raw laser micro-dissected HRS cell microarray data (Affymetrix GeneChip HG-U133 Plus 2.0 platform) from 29 HRS patient samples and 5 germinal center B-cell (GCB) controls was acquired from C. Steidl [152]. Microarray probes against the nine protein-coding genes with evidence of cancer-specific TE-initiated isoforms were manually extracted. The raw data was log2 transformed and the fold-change of each probe expression value was compared against the mean of the GCB controls.

The NanoString nCounter Elements (Seattle WA) platform was used for digital gene expression profiling on 100 ng of RNA. A custom-designed code set was used termed HL-LTR (Supplementary Table 4.5, 4.6) targeting 27 distinct isoforms of 14 genes with either canonical or non-canonical LTR- initiation sites shown to be activated or substantially up-regulated in cHL relative to normal

controls, and 6 housekeeping control genes. Housekeeping genes (*TBP, SDHA, WBP4, POLR1B, GUSB* and *TNFRSF8*) were selected to have (1) stable expression across 314 cHL patient biopsies (NanoString RHL800 panel [279]), as determined by the geNorm algorithm (implemented in NormqPCR R package [289]) and (2) relatively lower total RNA expression such that expression signals of rare transcripts are not inhibited.

NanoString platform data was normalized per manufacturer's recommendations [290]. Briefly, the mean of the negative control probes in each sample was subtracted from the raw counts. The housekeeping normalization factor was calculated by the sum of the housekeeping probe read counts per sample, divided by the geometric mean sum of the housekeeping probes across all samples. The read counts were then multiplied by the housekeeping normalization factor in each sample to yield a normalized probe expression value.

4.2.6 Statistical testing

Error bars shown are standard error of the mean, unless otherwise stated. Two-tailed Welch's t-test was performed to test for difference in the means with unequal variance using GraphPad Prism 5.0.3 for Windows (GraphPad software, La Jolla California USDA). Two-sided Student's t-test of microarray data was processed in R statistical language with a custom script.

4.3 Results and discussion

To identify instances of onco-exaptation in lymphoma, I screened HL cell lines, DLBCL patient samples and normal B-cell RNA-seq libraries. One candidate gene, which was also previously identified by a former post-doctoral fellow in the lab, was the proinflammatory transcription factor (TF) interferon regulatory factor 5 (*IRF5)*, which is recurrently up-regulated in HL derived cell lines. IRF5 belongs to a multi-member family of TFs responsible for inducing transcription of cytokines and chemokines in response to interferon signaling [291] but had not been implicated in

HL before. Together the bio-medical applicability of the *IRF5*, previously identified *CSF1R* and an additional set of newly characterized chimeric transcripts are explored.

4.3.1 TE-initiated transcripts are upregulated in lymphoma

To evaluate if there is an increase of TE-initiation events in the lymphoma transcriptome, *LIONS* analysis was run on 9 Hodgkin lymphoma (HL) cell lines, 3 primary mediastinal large B-cell lymphomas (PMLBL) cell lines, 66 diffuse large B-cell lymphoma (DLBCL) patient samples and 9 germinal center B-cell biopsy controls (Supplementary Table 2.1).

The total number of TE-initiated events are non-significantly increased in HL and PMLBL RNA-seq libraries relative to the B-cell controls (Figure 4.1A). However, when partitioned into TE-families, a specific and significant increase in LTR-initiated events in HL is evident (Figure 4.1Aii). This supports the hypothesis of transcriptional activation of TEs in lymphoma, specifically among LTRs.

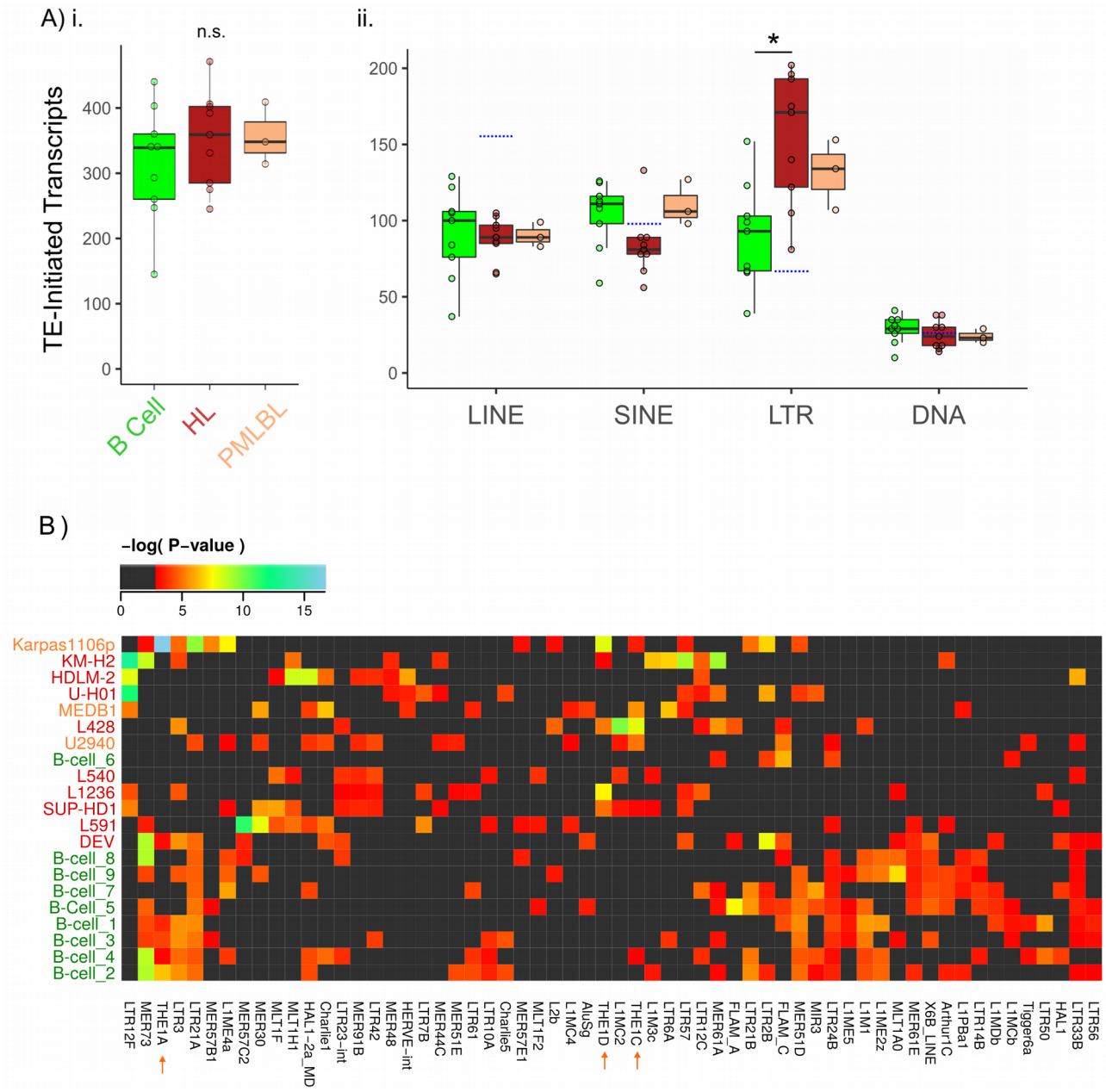


Figure 4.1: TE-initiated transcripts in Hodgkin Lymphoma

The **A) i.** Total and **ii.** class stratified *LIONS* detected TE-initiations in nine B-cell controls (green), nine Hodgkin lymphoma cell lines (red), and three Primary Mediastinal Large B-cell Lymphomas (peach). Blue dotted line shows the expected distribution of TE classes based on the relative number abundance of each TE class in the genome. **B)** Exact binomial test for enrichment of repeats relative to the expected input abundance. TEs that are enriched ($p < 0.05$) in at least any 2 libraries are included and used for clustering of the libraries.

In the comparison between HL cell lines and B-cells, of which there are an equal number of samples: the HL set contains 2411 distinct TE-initiation sites, with 395 (16.4%) being recurrent (present in 2+/9), of which 311 are specific (absent from 9 B-cell controls) (Supplementary Table 1). In the inverse analysis, the B-cell libraries contain only 1573 distinct TE-initiated transcripts, with 495 (31.5%) recurrent TE-initiations, and 372 are specific (absent from HL). Similar to CRC cells, HL shows a high proportion of unique sites, and surprisingly a lower amount of recurrent and specific sites. This is unexpected as HL cells are derived from B-cells and are expected to contain epigenetic information from their ancestor state, although these are HL cell lines which may have diverged substantially from their normal epigenetic state. This does give rise to an intriguing corollary of the assumptions regarding oncogenic TE-initiations: endogenous TE-initiations with tumor-suppressor function are recurrent and specific to normal-healthy tissue and absent from the malignant tissue (if and only if the normal tissue represents the epigenetic cell of origin for the malignancy).

The THE1 elements, which are of the ERV-MaLR class, have been postulated to be significantly enriched in HL, owing to an analysis originating from the fact that the oncogene *CSF1R* is ectopically driven by a THE1B element [292–294]. The THE1 elements are highly abundant in the human genome, on the order of 37,000 independent LTRs [2], which creates many opportunities for onco-exaptations to occur. Targeted LTR-initiation studies using RACE-seq have focused on these elements, and are largely based on L428, L1236 and KM-H2 cell lines [294]. The THE1 elements are not consistently enriched in cHL, PMLBL, or B-cells (Figure 4.1B). THE1A elements are moderately enriched in some B-cells. Across the cHL, the THE1D element was enriched in L428, L1236, KM-H2 and SUP-HD1, but also in the Karpas1106p cell line. Overall, no individual group of LTRs is enriched in cHL, and the emphasis of the role of THE1 elements [294] may in part be due to the focus on the cell lines chosen.

To identify transcripts that are of biological relevance to cancer biology, two simplifying assumptions regarding the distribution of events were made. *Assumption of recurrence*: TE-initiation events which promote oncogenesis will arise multiple independent times in different patients. This assumption removes one-off or rare TE-initiation events, focusing on the instances which arise at a higher rate, affecting more patients. *Assumption of specificity*: TE-initiation events that promote oncogenesis will arise in the lymphoma libraries and not in "normal" control libraries. This assumption removes TE-initiation cases which have undergone evolutionary (normal) exaptation for use in the transcriptome. It is reasonable that these two assumptions will be sufficient to identify oncogenic TE-initiation events, but the converse is not true, not all recurrent and specific TE-initiation events are necessarily oncogenic, they may also be a correlate of a hidden variable in the cancer. HL recurrent and specific transcripts are listed in Supplementary Table 4.1 and investigated in more detail in sections 4.3.2 and 4.3.3.

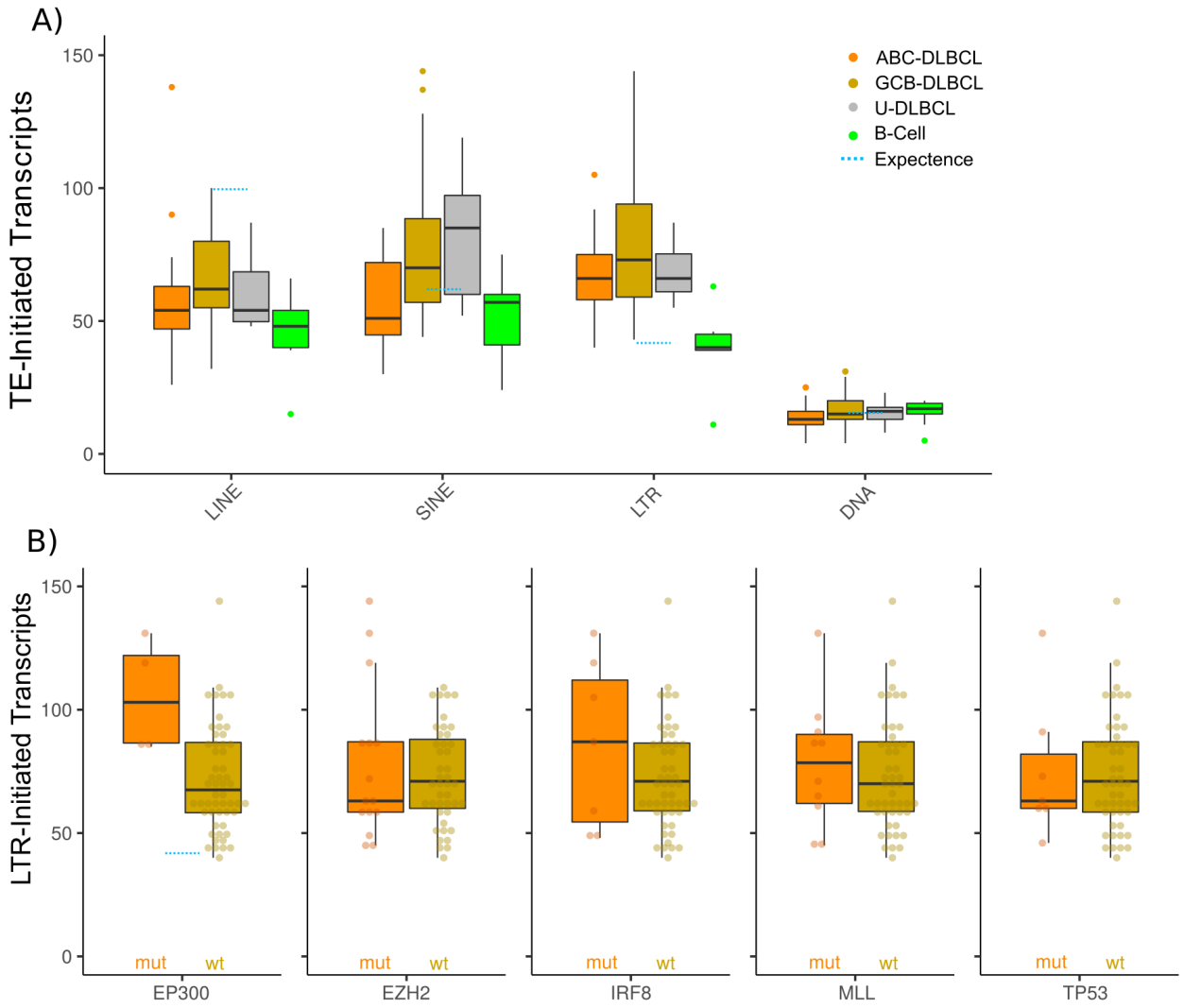


Figure 4.2: TE-initiated transcripts in Diffuse Large B-cell Lymphoma

Sixty-six DLBCL patient biopsy RNA-seq libraries were analyzed by LIONS. A) TE-initiated transcripts in activated B-cell (ABC), germinal B-cell (GCB), unclassified (U) DLBCL or B-cell controls. B) LTR-initiated transcripts, which show the most responsive TE-activation, sub-classified by the mutational status of the patient sample in key epigenetic regulators. There is no statistical difference (Welch's two sample t-test) between between mutational status.

I also analyzed primary patient data from a previous study in our laboratory of DLBCL data using a simpler method [230] and identified at least 97 chimeric transcripts [180] but that method was non-quantitative with a significant false positive rate that did not allow in depth statistical analysis. *LIONS* analysis identified 5216 TE-initiated transcripts in the DLBCL data-set, of which

1846 (35.4%) occurred in greater than two libraries. Hypothesizing that mutation of key epigenetic regulators [295] in DLBCL may correlate with TE-derepression, the DLBCL patients were intersected by mutational status and TE-activity compared. After correction for multiple testing, there was no difference between germinal center B-cell (GCB) and activated B-cell (ABC) subtypes of DLBCL. Also surprisingly, the mutation status of *EP300, EZH2, IRF8, MLL and TP53* all were not associated with an increase in TE- or LTR-initiated transcription (Figure 4.2). In light of this, it appears that single mutational events resulting in epigenetic perturbation do not globally result in LTR-derepression. This is in contrast to the fibroblast 5-aza treatment model (Chapter 3) which shows an immediate response upon epigenomic disruption.

4.3.2 The onco-exaptation of *IRF5*

To screen for TE-gene chimeric transcripts in HL, paired-end RNA-seq reads were analyzed in 9 HL, 3 primary mediastinal large B-cell Lymphoma (PBMCL) derived cell lines [241] and 9 normal CD77+ centroblast B-cell controls [179] (Table 2.1). The screen identified a TE-initiated transcript from a LOR1a LTR element upstream of *IRF5* which was present in 7/9 HL lines (not detected in UH-O1 and the NPL-HL line, DEV), 1/3 PMBCL (MEDB1) and 0/9 B-cell samples (Figures 4.3 and Supplementary Figure 4.1). [281][282] Enticingly, during the course of my thesis an independent study of genome-wide DNase hypersensitivity data by Kreher et al., identified *IRF5* as being a pivotal TF upregulated specifically in HL cells and crucial for their survival. Further, IRF5 cooperates with NF-κB as a central regulator of the HL transcriptome [297]. Here I show that transcriptional activation of a normally dormant LTR plays a significant role in the upregulation of IRF5 in HL. Hence, the HL-associated deregulation of at least two genes with major roles in this disease, *CSF1R* and *IRF5*, is mediated through the awakening of ancient LTR promoters. The transcription start site within the LOR1a element was validated by 5’ RACE (Figure 4.3B). To

determine the tissue specificity of chimeric *IRF5*, I inspected ENCODE RNA-seq data from 17 cell lines and 31 normal primary tissues and no evidence for LOR1a-*IRF5* chimera was found except for the three EBV-transformed B-cell lines GM12878, GM12891 and GM12892 (Supplementary Table 4.3). The absence of IRF5 chimera in primary tissues, particularly lymphocytes and leukocytes, suggests that the LOR1a LTR transcriptional activity is a transformed B-cell specific and recurrently occurring phenomenon. Recently, EBV-induced transformation was shown to induce upregulation of LTR-initiated transcripts, consistent with my observations [281]. In fact, several promoters for *IRF5* have been described in normal cells [282], while this LTR has not previously been characterized as a promoter. The chimeric transcript contains the complete open reading frame for *IRF5*, which begins in native exon 2, and full-length chimeric *IRF5* cDNA could be PCR amplified (Figure 4.3, Supplementary Figure 4.1).

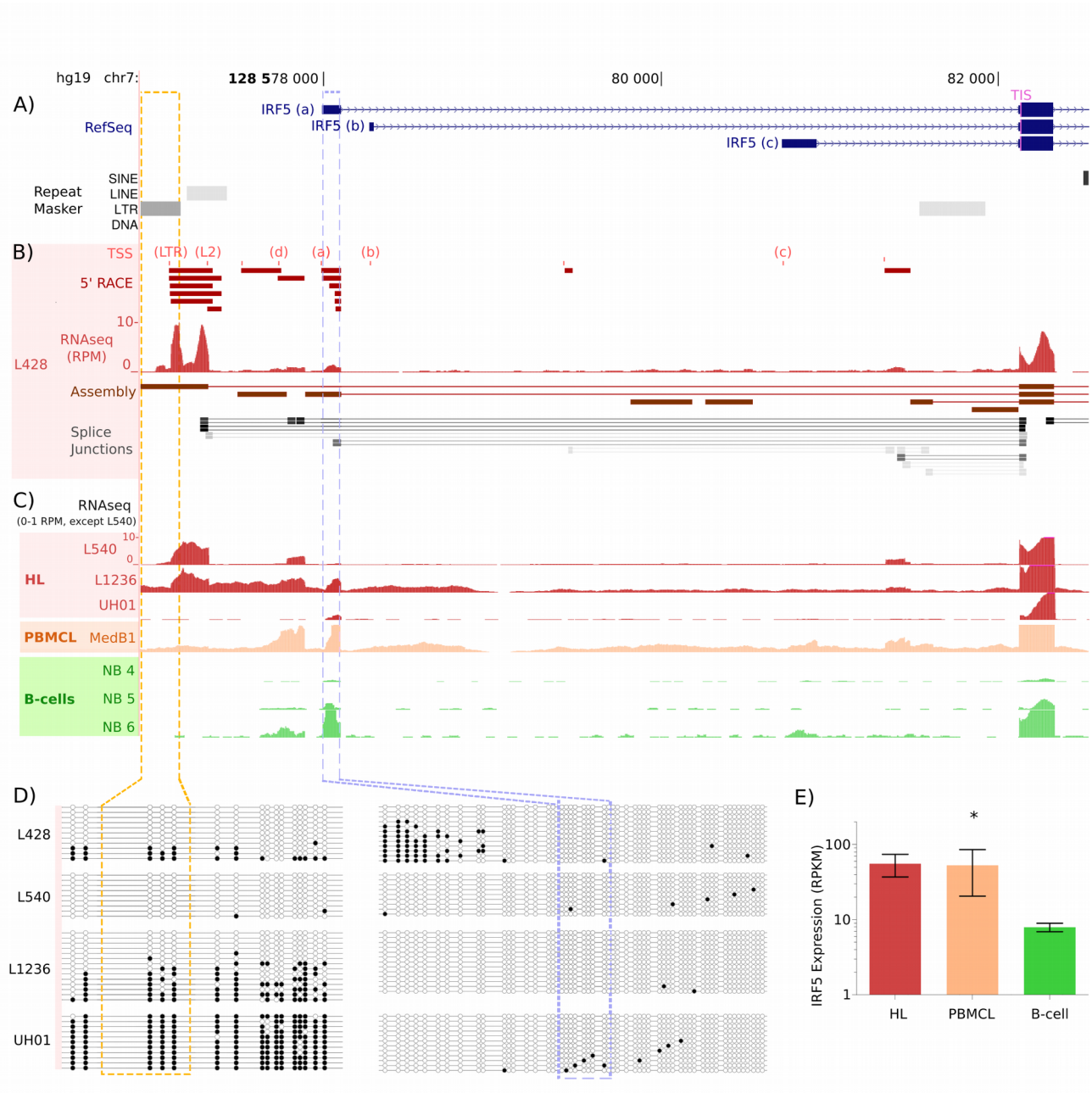


Figure 4.3: A LOR1a LTR element drives IRF5 expression in Hodgkin lymphoma

A) *A UCSC genome browser view of the 5' end of RefSeq annotated IRF5, RepeatMasker defined transposable elements and IRF5 transcription start sites (TSS) for native isoforms a-d [352] and LTR, L2 isoforms described in this thesis. The IRF5 translation initiation site (TIS) begins in the native exon 2.* ***B)*** *Hodgkin Lymphoma (HL) L428 cell line RNA-seq coverage plot of uniquely mapped reads in Reads Per Million (RPM) shows expression of upstream exons initiating within a LOR1a LTR, relative of the native IRF5* transcripts and u*nique first exons in L428 determined by 5` RACE and ab initio RNA assembly tracks. Splice junctions are shaded by supporting reads from one, pale gray, to >=20, black.* ***C)*** *Representative HL RNA-seq coverage scaled from 0-10 RPM for L540 and 0-1 RPM (L1236 and UH01) showing a range of LTR promoter usage from high in L540 to absent in UH01. Representative PBMCL line, MedB1 (orange), and normal B-cell transcriptomes (green) predominantly transcribe IRF5 from the native isoform a and d promoters but lack transcription from the LTR. Complete...*

Text 5: Figure 4.3 Continued

… panel in Supplementary Figure 4.1. ***D)*** *Bisulphite sequencing of the LTR and native promoter regions with open circles showing unmethylated CpGs and solid circles methylated CpG sites. Cell lines with active LTRs are hypomethylated while UH01 which uses the native promoter is hypermethylated.* ***E)*** *Total expression of IRF5 in HL (n = 9), PBMCL (n=3) and B-cell (n = 9) RNA-seq libraries calculated as reads per kilobase per million reads (RPKM). Error bars are the standard error of the mean. Two-tailed Welch's t-test was performed to test for difference in the means with unequal variance with p-values equal to 0.0332 (*) between HL and B-cells.*

To assess the epigenetic state of the LTR element between chimera positive and negative HL I investigated the methylation status of both the native and LTR promoters. In chimera positive L428, L540 and L1236 cells the LOR1a LTR exists in a hypomethylated state while in the chimera negative UHO1 cells, the LTR was hypermethylated (Figure 4.3D). The primary native promoter (start site “a”) exists within a CpG island and is unmethylated regardless of activity (Figure 4.3). LTR derepression further correlates well with expression of the LOR1a-*IRF5* isoform relative to the native promoter isoform, and a proportional increase in the total IRF5 protein (Figure 4.4). By mapping the available DNase I hypersensitivity data [297] of HL and non-HL cell lines, we observed that the hypomethylated LTR in L1236, L428 and L591 cells was within a DNase I hypersensitive region, while it was not in the non-HL lines Namalwa and Reh (Figure 4.5A). Together, the absence of DNA methylation and the open chromatin state suggests that this locus would be accessible to transcription factors and the transcriptional initiation machinery.

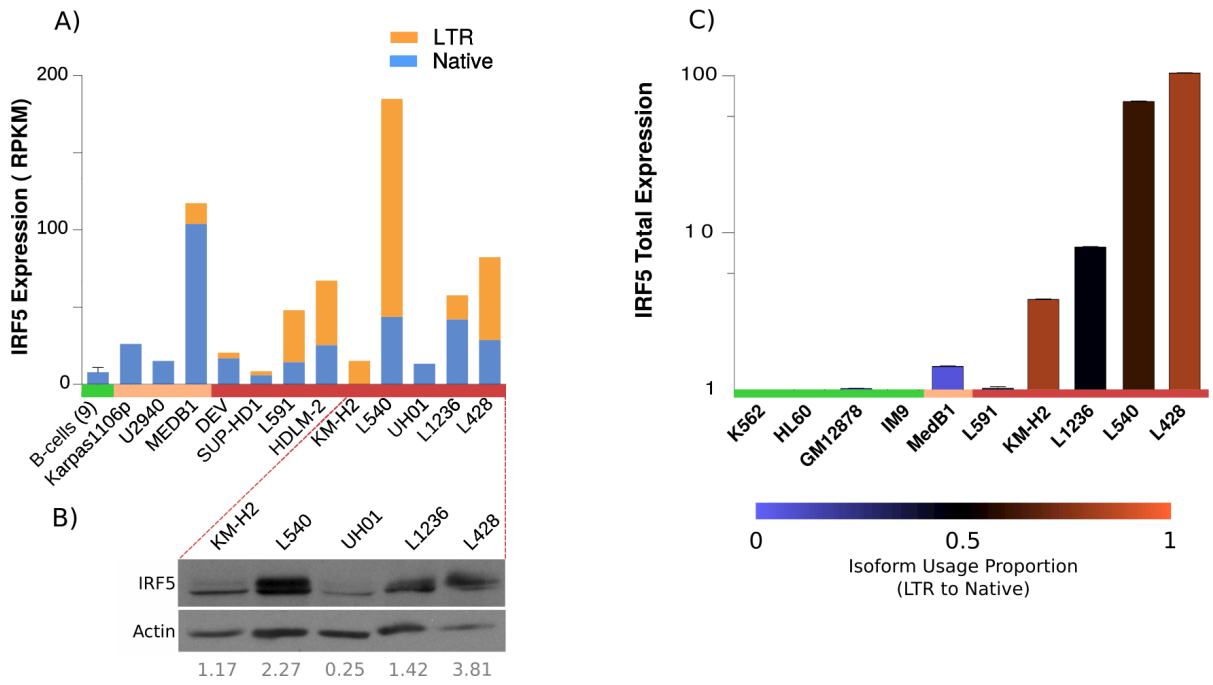


Figure 4.4: LTR contribution to IRF5 mRNA levels and total protein

A) *Promoter contribution of LTR and native first exons to IRF5 was calculated comparing the RPKM across all LTR- or Native-promoter splice junctions segments* ***B)*** *Western blot against IRF5 and beta-Actin of HL cell lysates (KM-H2, L540, UH01, L1236, L428) and IRF5:Actin protein band intensity quantification is shown below each lane.* ***C)*** *Quantitative RT-PCR was performed on cDNA from different HL lines to assess the relative expression level of native 'a-isoform' IRF5 and LOR1a-IRF5. Total IRF5 levels were measured using primers targeting exons 2 and 3 (downstream of both promoters converging). The relative promoter activity was measured as a ratio of LTR-specific to Native-specific transcription.*

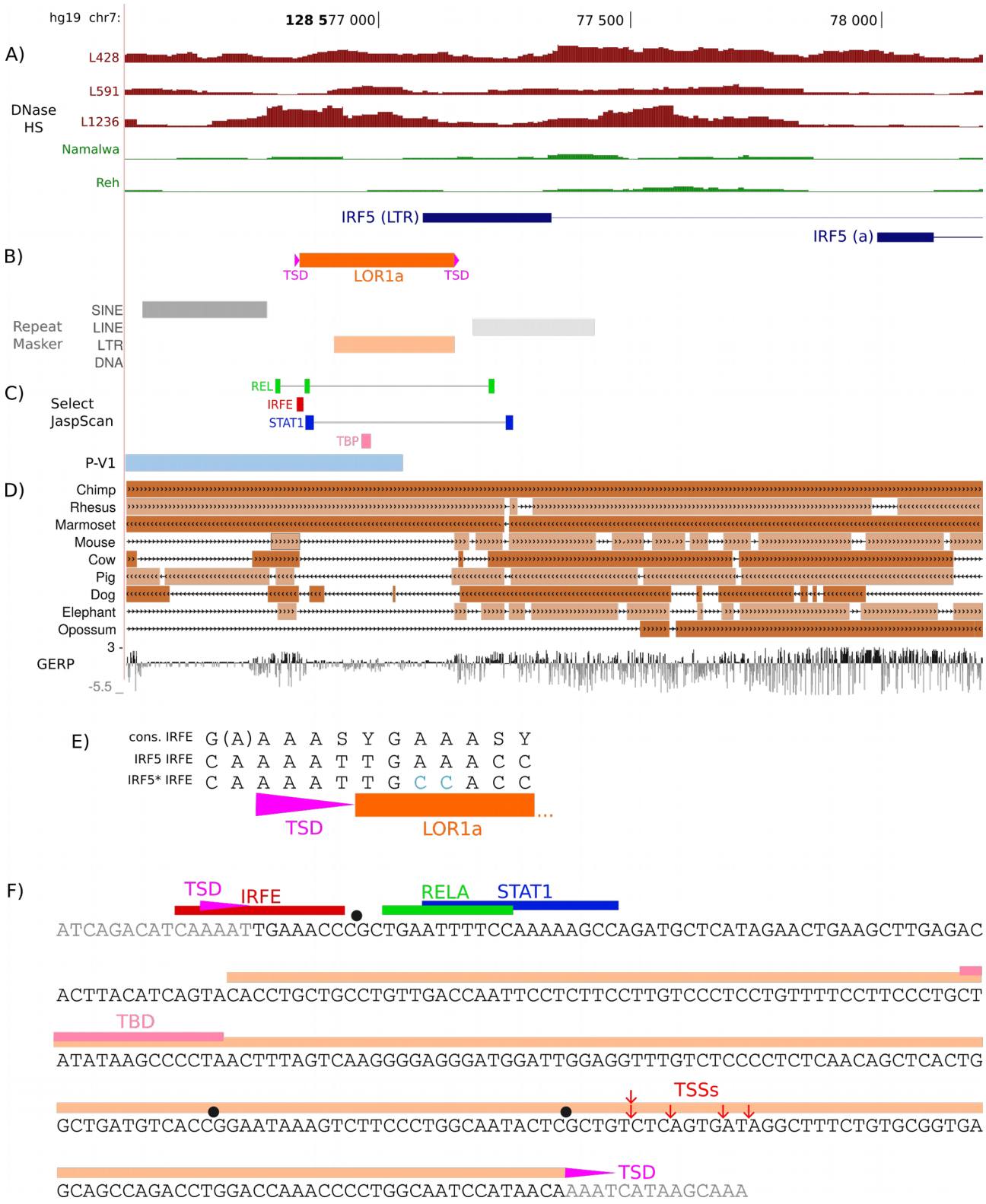


Figure 4.5: Features of the LOR1a LTR genomic region

Text 6: Figure 4.5 Continued.

A) *DNase hypersensitivity tracks [297] for three HL and two non-HL cell lines show open chromatin conformations over the LOR1a LTR in lines expressing chimeric IRF5. The 5` ends of the LTR-initiated transcript and the native "a" transcript, are shown in dark blue below the DNase I tracks.* ***B)*** *The inferred complete LOR1a LTR, shown as an orange bar above the Repeatmasker track, was identified by the tandem site duplication (TSD, magenta triangle) and homology of the upstream region to different LOR1a elements found in hg19 identified via BLAST alignment. The LOR1a extends past the RepeatMasker annotation.* ***C)*** *Select JaspScan [300] motifs identified in the LOR1a include REL, IRF and STAT and TATA-binding Protein (TBP) binding sites. The 'P-V1' promoter region analyzed by Mancl et. al [299] is shown in light blue.* ***D)*** *Multiple species alignments [353] and Genomic Evolutionary Rate Profiling (GERP) score [287] show that the LOR1a retrotransposition occurred in a common primate ancestor.* ***E)*** *The consensus interferon regulatory factor binding element (IRFE), the sequence found in the human genome upstream of (IRF5) and the inactive/mutated sequence (IRF5*) previously identified [299] are shown aligned to the "AAAT" TSD sequence and beginning of the LOR1a LTR "TGAAACC".* ***F)*** *Nucleotide sequence of the IRF5-LOR1a LTR element in black with flanking sequence in gray. The RepeatMasker annotation (over-lined with light orange) for the LOR1a misses the 5' end of the LTR which was identified by alignment and the characteristic target site duplication (TSD), "AAAT". Transcription start site (TSS, red arrows) defined by 5' RACE clones from L428 and CpG sites are shown with black circles.*

Little is known about the LOR1a family of LTRs beyond an entry in the repetitive sequence database, Repbase, that reports a consensus LTR sequence of 497 bp (Figure 4.5B). The LOR1a LTR locus upstream of IRF5 is only 239 bp and the Repeatmasker annotation suggests it is missing the 5` end. To investigate the LTR structure further, I retrieved the non-TE 134 bp immediately 5` of the annotated LTR and looked for related sequences throughout the genome. Alignment of this upstream region to the hg19 human genome identified 34 homologous sequences of 69 bp upstream of other regions annotated as LOR1a (Supplementary Table 4.4), suggesting that the full LTR of this distinct subfamily is longer than annotated. Indeed, by examining the termini of these extended LTRs, I was able to identify putative 4 bp target site duplications (TSD), that are created upon integration of retroviruses [298] and therefore deduce the full length of these LOR1a subtypes, which is 308 bp for the copy upstream of *IRF5* (Figure 4.5D-F, Supplementary Table 4.4).

Evolutionary sequence comparisons indicate this LTR copy integrated at least 45-50 Ma, since it is present in both New and Old World primates but is absent in non-primates (Figure 4.6D).

Although no mention was made of the LTR, Mancl et al. previously investigated the promoter activity of a region called “P-V1” surrounding this LTR (Figure 4.6C) and identified within it a critical interferon regulatory factor binding element (IRFE) that controls promoter activity in a luciferase reporter assay in response to various IRFs, in particular IRF5 itself [299]. I identified the same IRFE using JaspScan [300] within this region and, intriguingly, found it to be located directly at the boundary of the LOR1a and the TSD, such that the IRFE site contains the TSD and first few bases of the LTR (Figures 4.5E,F). This transcription factor binding site was therefore created serendipitously millions of years ago when the LOR1a element retrotransposed. Hence, the inherent core promoter motifs of an LTR plus the formation of an IRFE site unique to this integration event have combined to create this active promoter in HL.

In conclusion for this section, I have shown that the LOR1a LTR upstream of *IRF5*, which is dormant in normal tissues, has been re-purposed in HL, resulting in LTR promoter activation and associating with overexpression of *IRF5*. While IRF5 is oncogenic in HL [297], the necessity and sufficiency of specifically the LOR1a driven *IRF5* transcript to oncogenesis requires experimental validation via isoform specific knockdown of LOR1a-*IRF5* or knockout of the LTR in HL cells. This onco-exaptation occurs recurrently in multiple independent HL lines suggesting overexpression of IRF5 may be selected for and the LOR1a IRFE site provides an exploitable genetic circuit for this. *IRF5*, along with *CSF1R* [150] and *FABP7* [180], are the best characterized examples of onco-exaptation of LTRs in lymphoma, but this is likely to be a broadly occurring phenomenon in oncogenesis (Chapter 1). Taken together, these studies establish that cancer-specific transcription driven by activated LTRs or other TEs, namely onco-exaptation, is a distinct and

under-investigated mechanism for oncogene activation, with a unique etiology and possibly, a novel diagnostic or prognostic indicator in patients.

4.3.3 Biomarker potential of TEs in Hodgkin lymphoma

The rare nature of HRS cells makes them difficult to interrogate directly, so most prognostic indicators have focused on markers in the immune or stromal cell microenvironment [278]. In a laser micro-dissected micro-array analysis specific to the HRS cells, *CSF1R* was expressed in 48% of cHL cases (which is dependent on the THE1B LTR element [150]), and correlated with poor progression-free and overall patient survival [152]. Here I investigated whether the LTR-mediated mechanism of oncogene activation of *CSF1R* and *IRF5*, and the recurrent and specific TE-initiated transcripts I identified using *LIONS* in cHL, could be exploited to develop a unique biomarker assay, specific to the cancer-specific TE-initiated isoforms.

Single molecule RNA hybridization technologies such as NanoString can discriminate transcript isoforms when the detection probes are designed against the exon-exon junctions unique to a transcript isoform, in fact this has been already applied for the detection of the LTR-initiated *ALK* isoform [161,301]. In the example of *CSF1R*, the native isoform is highly expressed in the myeloid lineage and therefore is highly expressed in cHL tissue from the tumor-associated macrophages [153,276], but, *CSF1R* is prognostic when it is expressed from the HRS cells [152,153]. It is possible to design NanoString probes specific against the splice junctions of the THE1B-CSF1R isoform, and therefore measure HRS cell *CSF1R* expression from a complex tissue biopsy. Such reasoning can be applied to an entire set of TE-initiated gene transcripts, allowing the development of a cHL biomarker panel based on the molecular state of the cancer itself and not simply its microenvironment.

The HL-LTR panel (Table 4.1) was rationally designed based on a cross-reference of cHL cell line, B-cell, and ENCODE, and normal tissue RNA-seq and CAGE data. In addition to the multiple

TE-initiated isoforms of *CSF1R* (a known prognostic factor), the CSF1R ligand *CSF1*, and *IRF5* (a known cHL oncogene), and several TE-initiated transcripts from the *LIONS* analysis (Chapter 3) were selected. In total the HL-LTR panel includes 37 probes targeting both TE-initiated and native promoter initiated (and/or total) transcripts and six housekeeping genes (Supplementary Table 4.5). The protein-coding genes *CSF1R*, *IRF5*, *VASH2*, *FHAD1*, *CSF1*, *RALB*, *UNC13C*, *DHRS2*, *IL1R2* and *ZNF281* are represented. The panel also includes the non-coding transcripts *KIRREL3-AS1*, *AFAP1-AS1*, *ZNF281-AS1*, in addition to uncharacterized lncRNA nc*CSF1* (non-coding RNA upstream of *CSF1)*, and hlnc1 (Hodgkin lymphoma specific lncRNA 1).

Several of these genes can be reasonably hypothesized to be onco-exaptation events, based on what is known about the gene function. These include:

VASH2: Vasohibin 2, an angiogenesis inhibitor, is overexpressed in various solid cancers, where it has been implicated in cancer progression, inducing angiogenesis, tumor growth and epithelial-mesenchymal transition (EMT), at least partly by activating TGF-beta signaling [302–304]. Nuclear VASH2 has also been reported to induce cell cycle progression and proliferation [305]. VASH2 has not been studied in any blood cancers but the fact that it is recurrently and specifically expressed from an LTR promoter in HL cell lines (Figure 4.6) and is upregulated in primary HRS cells (Figure 4.7) makes it an intriguing target.

FHAD1: Forkhead Associated Phosphopeptide Binding Domain 1 is expressed in lung, testis and fallopian tubes but little is known about its function or potential role in cancer, although hypomethylation of the *FHAD1* promoter is associated with poor patient outcome in prostate cancer [306]. I have chosen it because of high recurrence of LTR-driven FHAD1 in the cHL cell lines with very little evidence of LTR usage in other cell types (Figure 4.6, Table 4.1).

Gene	Gene function	TE	TE coordinates	RNAseq cHL	B-cells	Fetal	Adult	Cell Lines	Microarray Fold (HRS / B-cells)
CSF1R	Receptor Tyrosine Kinase	THE1B : ERVL-MaLR	chr5:149472016-149472372	**5**	-	-	-	1	2.289
IRF5	Transcription factor	LOR1a : ERV1	chr7:128576913-128577151	**7**	-	-	HSC	3	2.131
FHAD1	Forkhead associated domain; Binds pSer, pThr, pTyr	MLT1K : ERVL-MaLR	chr1:15562769-15563305	**6**	-	-	-	3	1.409
VASH2	Angiogenic Vasohibin And pro-cycling	MLT2B2 : ERVL	chr1:213104237-213104759	**4**	-		-	4	7.458
DHRS2	Dehydrogenase	LTR12D : ERV1	chr14:24104836-24105861	**8**	-	2	12	11	256.885
IL1R2	Decoy Cytokine Receptor	MLT1H1 : ERVL-MaLR	chr2:102614909-102615507	**8**	-	1	PBMC, Thy. + 6	5	52.116
RALB	RAS-like protooncogene B, GTPase	THE1C : ERVL-MaLR	chr2:121013952-121014311	**7**	-	-	-	1	2.198
ZNF281-AS1	lncRNA, vertabrate conserved, Stem cell expressed	MER21B : ERVL	chr1:200380599-200381432	**6**	-	9	HSC + 3	3	na
ZNF281	Transcription factor, Regulates pluripotency	MER5B : DNA-hAT-Charlie	chr1:200452608-200452783	**6**	-	-	1	4	0.912
NC-CSF1	lncRNA upstream of CSF1	LTR8 : ERV1	chr1:110374684-110374922	**4**	-	2	HSC	2	na
CSF1	Ligand for CSF1R RTK	LTR8 : ERV1	chr1:110374684-110374922	**1**	-	-	-	1	2.105
KIRREL-AS1	lncRNA, intronic Antisense to Kirrel3	MSTA : ERVL-MaLR	chr11:126517838-126518201	**6**	-	-	-	4	na
HLNC1	lncRNA, high expression, HL-specific	THE1C : ERVL-MaLR	chr2:8071936-8072307	**5**	-	-	-	0	na
AFAP1-AS1	lncRNA antisense to AFAP1, exons overlap	THE1A : ERVL-MaLR	chr4:7755611-7755963	**3**	-	-	3	4	na
UNC-13C	UNC13 homolog; brain Expressed; in-frame	MER73 : ERVL	chr15:54875841-54876417	**3**	-	-	-	4	1.813
			(hg19)	(/ 8)	(/ 9)	(/ 11)	(/ 36)	(/ 20)	

Table 4.1: HL-LTR target gene panel

Protein coding and non-coding targets selected for the HL-LTR panel, with specificity and/or substantial upregulation in cHL and/or cancer cell lines. From 36 adult tissues; there is RNA-seq evidence for alternative isoform expression in the hematopoietic lineages in hematopoietic stem cells (HSC), peripheral blood mononuclear cells (PMBC) or thymus (Thy) samples.

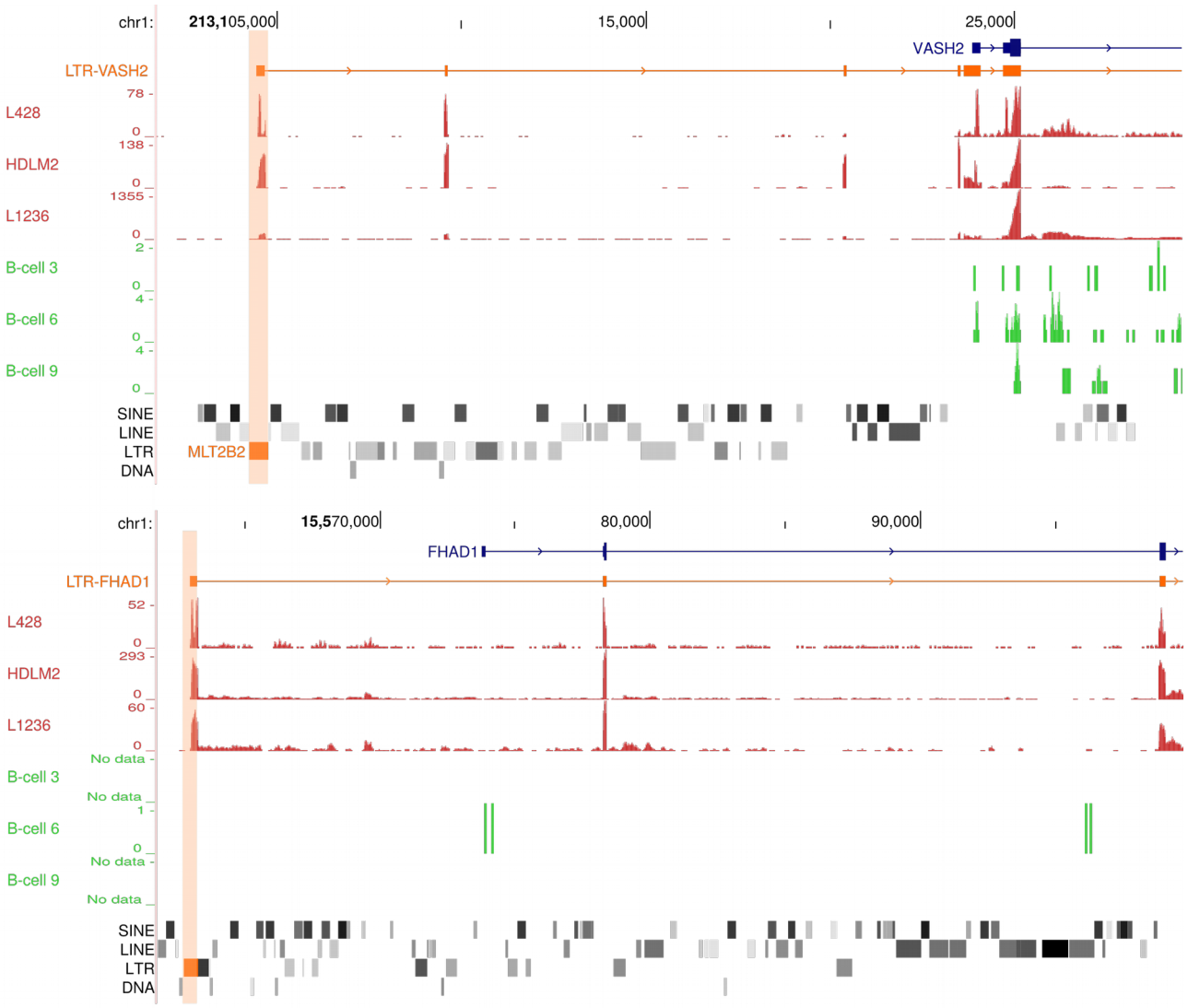


Figure 4.6: LTR-initiated transcripts of VASH2 and FHAD1

UCSC Genome browser screen shot showing the coverage over the LTR-initiated transcripts in the L428, HDLM2 and L1236 cHL cell line and B-cell control RNA-seq. A) The MLT2B2-*VASH2* transcript splicing upstream of the native first exon of *VASH2*. B) The MLKT1K-*FHAD1* transcript splices from the alternative first LTR exon directly into the native second exon containing the CDS start site, and bypassing the canonical exon 1.

RALB: RAS-like protooncogene B, is a small GTPase activated immediately downstream of RAS. *RALB* has been implicated in promoting cancer-cell migration and invasiveness both *in vitro* and *in vivo* [307–309]. In a model of AML, *RALB* activation was capable of alleviate tumors of *NRAS[G12V]* addiction, demonstrating that a major effector pathway of common *NRAS* mutation is

mediated through the RALB signaling, and not necessarily MAPK/PI3K signaling [310]. The recent research into *RALB* has led to the exploration of this GTPase as a therapeutic target for RAS mutated cancers [310,311].

To test if the HL-specific TE-initiated transcripts identified from the cell line RNA-seq and listed in Table 4.1 were upregulated in primary cHL patients, HRS laser micro-dissected microarray gene expression data for 29 HRS patients was compared to 5 GCB controls [152]. The microarray design includes probes predominantly for protein-coding genes and not lncRNA, thus from the 12 included genes (across 32 probes), 8 genes (11 probes) are significantly upregulated (two-tailed Students T-test, $p < 0.05$) in cHL patients (Table 4.1 and Figure 4.7). The microarray probes predominately target 3' UTR and therefore would assay both native and TE-initiated isoforms, in addition it is not anticipated that *every* cHL sample is positive for every TE-initiated transcript in the HL-LTR panel, but these preliminary results support that this assay can be developed in future work as an informative biomarker for sub-set of patients with HL.

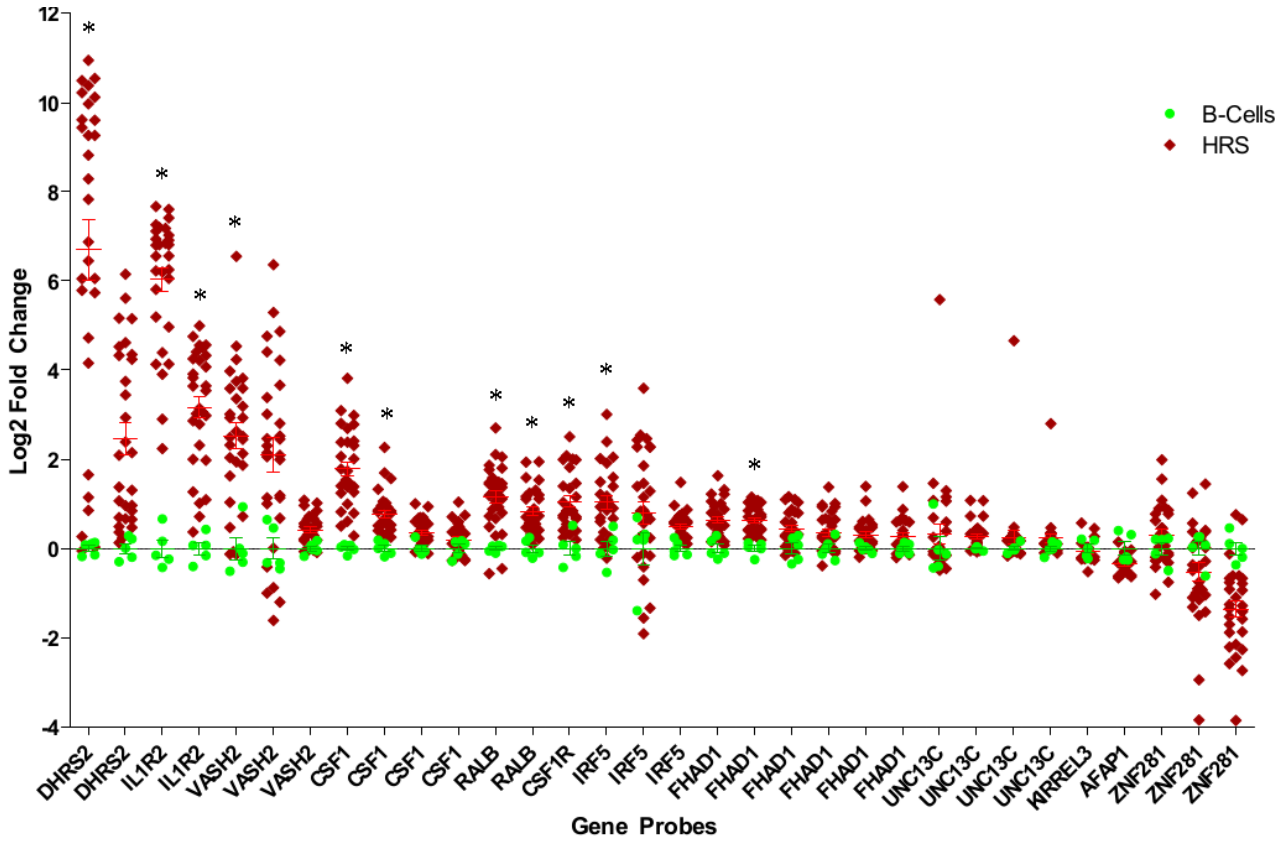


Figure 4.7: HL-LTR panel gene expression in micro-dissected HRS vs. GCB controls

The log2 fold-change gene expression profile of twelve protein-coding genes included in the HL-LTR biomarker panel. Gene expression from 29 cHL patients (red diamond) where HRS cells were isolated by micro-dissection was compared to 5 germinal center B-cell (GCB) controls (green circles). A star denotes a significant difference (two sided Student's t-test, $p < 0.05$) between biological groups.

In a pilot experiment to measure the relative expression of HL-LTR transcripts, RNA from five cHL cell lines and four non-HL cell lines was hybridized with probes and measured with the NanoString nCounter system (Figure 4.8). As designed, cHL cells on average show higher expression of all target genes. Some non-cHL cell lines also show expression of one or more target transcripts, such as *DHRS2* in HepG2 hepatocellular carcinoma cell line and HEK293T embryonic kidney cell line (note: the native *DHRS2* promoter is an LTR), while other transcripts such as *hlnc1* show very high specificity for cHL. Given that HRS cells typically make up between 0.1-1%, but can reach 10% of the cells in a lymph node biopsy [274,312] of the cells of biopsy tissue, and

assuming an equal RNA content between HRS cells and niche cells, transcripts exceeding a normalized expression of ~800 (for HRS 1%) to 8000 (for 0.1% HRS) would be detected from HRS cells (using a detection limit of 8 counts from the negative control probe-set). The detection limit can further be lowered with assay optimization, currently target genes account for only 46.5% (+-20.7%) of the total NanoString count in the assay. Reducing the internal positive control probe abundance by a factor of ten would boost target signals by ~2.05 fold.

The rational design of the HL-LTR panel and proof-of-concept experiment demonstrates that exploiting TE/LTR-initiated transcripts for a future diagnostic or prognostic assay of complex tissues is possible but requires further optimization for NanoString sensitivity. Applying this or a similarly designed panel to a cohort of cHL patient biopsies would accomplish two aims. First, the presence/absence of each LTR isoform can form the basis of a patient classifier in a similar method to the rhl30 panel based on micro-environment gene expression [279], and knowing that *CSF1R* expression alone is predictive of poor patient outcome [152], the HL-LTR panel would add additional information upon which outcome-predictive models can be trained. Second, access to samples at progressive time-points of cHL would test the hypothesis that TE-initiated transcripts are acquired progressively in the course of cancer evolution. Combining the prognostic prediction weighting (assuming this proxies for positive selection) of each isoform in the panel, with its rate of acquisition would directly test if the HL-LTR set of cHL specific and recurrent transcripts are true onco-exaptation events. Beyond the biomarker utility of LTR-driven transcripts, it is possible that personalized therapies targeting the RNA of the HL-specific isoforms could be developed in the future to increase patient cure rates while decreasing toxicity of therapies to normal cells.

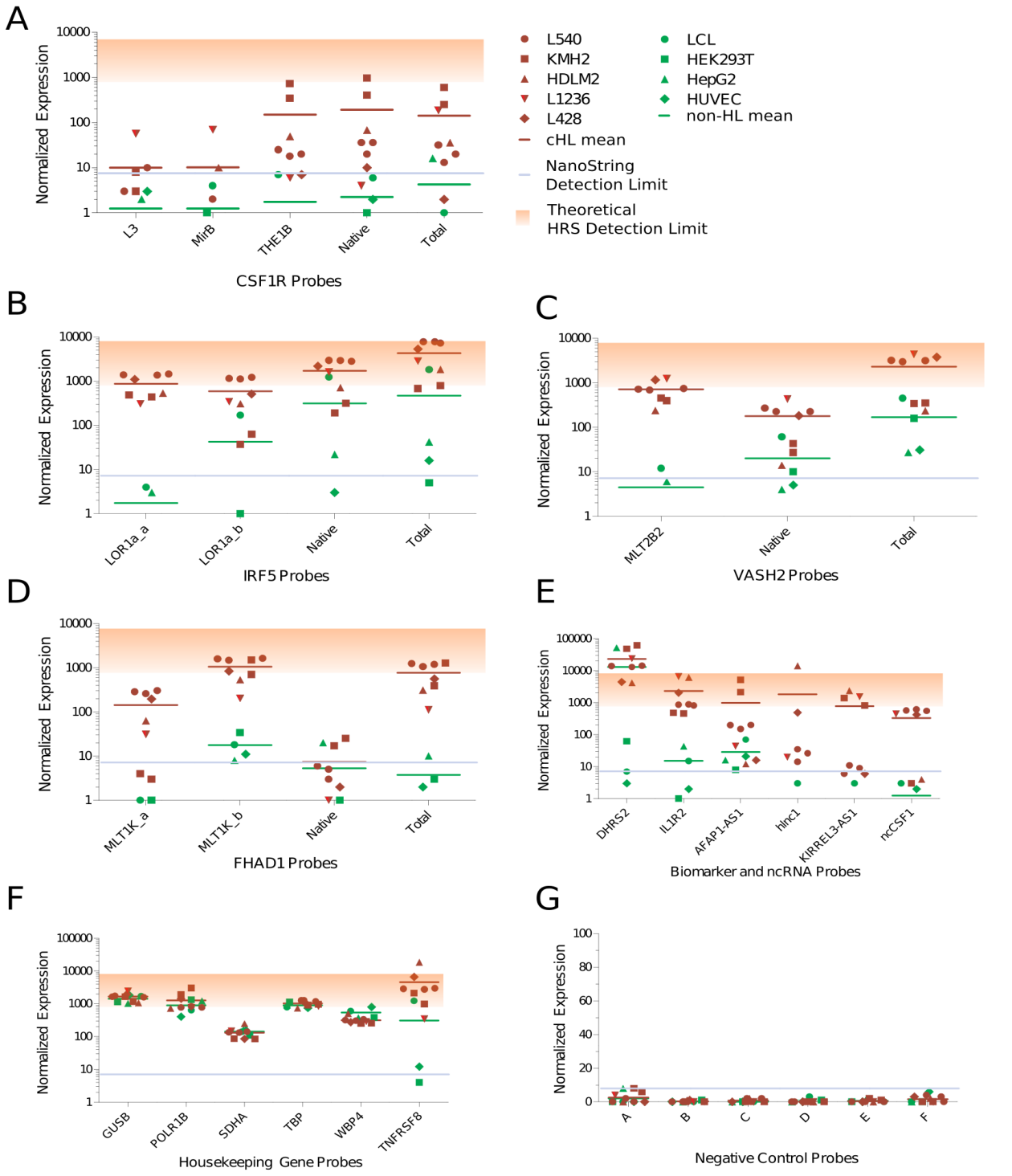


Figure 4.8: HL-LTR pilot experiment

Digital count expression values for NanoString HL-LTR assay in 5 cHL cell lines and 4 non-HL cell lines. TE-initiated targeting probes show a high expression specificity for cHL lines. Zero values are not drawn on the log-scaled graph. Red horizontal bar indicates mean of all cHL samples, green horizontal bar is mean of non-cHL samples. Pale blue line is NanoString detection limit determined by the maximum of the internal negative control probes (G). ...

Text 7: Figure 4.8 Continued

… The HRS-detection limit is a lower-bound for each probe under the presupposition that the cHL-specific isoform is below the NanoString detection limit in the non-HRS cells of the niche. Orange highlighted region is the upper (0.1% HRS cells) and lower bound (1% HRS cells) on the theoretical detection limit of the probe in HRS cells from a complex sample. A-D are probes targeting various isoforms of a single gene, and panel E are the probes for which only a the LTR-initiated isoform is is known, and therefore total gene expression was measured with a single probe.

4.4 Conclusions

Examples of LTR/TE onco-exaptation have expanded as sequencing techniques and the interest in this area of research developed. The activity of the LOR1a LTR upstream of *IRF5* in cHL is an important prototype of such oncogene activation as it demonstrates proto-oncogene over-expression above the physiological limitations of a native promoter. In contrast to THE1B-*CSF1R* in which the native promoter shows zero expression of the gene in the precursor and cancer cells, LOR1a-*IRF5* acts in addition to the native promoter.

As the evidence for distinct onco-exaptation events grows, novel translational applications of these findings need to be explored. Taking advantage of distinct RNA sequences arising at TE/LTR and genic splice junctions as a biomarker of onco-exaptation is one immediately apparent application of this research. The HL-LTR assay based on the NanoString platform is a proof-of-concept of such an application, designed to specifically address the molecular problem of proving rare cancer cells within highly heterogeneous samples.

Regardless of the underlying mechanism, onco-exaptation offers a tantalizing opportunity to model evolutionary exaptation. Specifically, questions such as “How do TEs influence the rate of transcriptional/regulatory change?” can be tested in cell culture experiments. As more studies that focus on regulatory aberrations in cancer are performed in the coming years, I predict that this phenomenon will become increasingly recognized as a significant force shaping transcriptional

innovation in cancer. Moreover, studying such events will provide insight into how TEs have contributed to reshaping transcriptional patterns during species evolution.

Chapter 5: Discussion and models

Prior to the discovery that DNA constitutes the physical basis of hereditary information and the double-helical structure, the geneticist Thomas Morgan defined a gene abstractly as a unit of heredity that when mutated results in a phenotype [313]. Thus, once the specific combination of DNA nucleotides was understood to form the basis of genetic material, Morgan's abstract gene became irreversibly linked to the information stored by the physical DNA sequence. This immediately gives rise to a glaring inconsistency: if DNA stores all hereditary information in the cell, then in a multi-cellular organism, cells giving rise to a differentiated tissue must transmit this 'cell-type' information through a change of its DNA. Early DNA-hybridization experiments quickly refuted this idea and established that complex organisms have the same DNA content across tissues and there exists hereditary information transmission outside of DNA, or *epi*-genetics [314–316].

Epigenetics is an oft misused term. The close relationship between epigenetically encoded information and DNA methylation, and chemical modification of histone tails may lead one to believe that epigenetics is the study of DNA methylation and chromatin. In fact, epigenetics is a much broader area of study, encompassing all non-DNA based hereditary information.

Broadly speaking, the prototypical epigenetic state of a transposable element in the human genome is repression. TEs are characterized by closed-chromatin, marked by proximal DNA methylation [317,318] and the repressive histone tail modification, H3K9me3 [319,320] and H3K27me3 [321]. TE repression is a necessary host defense mechanism to suppress the ectopic regulatory, transcriptional, and transpositional activities of these elements, which otherwise would perturb developmental regulation [318]. Murine studies have shown that the repression of TEs is established early in embryonic development [317] and persists (epigenetically) throughout the

animal's life. The repression of TEs appears to degrade over the lifespan of an organism, through a process of epi-mutation, not unlike time/cell division associated mutational accumulation [322].

The global DNA methylation of TEs, which is a proxy for repression, decreases with age [323,324], although it remains to be determined if this age-related TE demethylation is associated with regulatory or transcriptional activation of the TEs. There is evidence that an age-dependent increase in TE transposition occurs in somatic tissues, in particular LINE and Alu element accumulation [325,326].

Cancer is an age-related disease which arises due to mutational and epimutational processes [108]. The classical understanding of cancer is that DNA mutations lead to activation of proto-oncogenes and the suppression of tumor suppressor genes. In recent years this view has been expanded to include epigenetic variation as having a causative role in oncogenesis [327].

A primary line of evidence that epigenetic perturbation is causative in oncogenesis is that the genetic mutation of epigenetic regulators is common across multiple cancers [111]. For example, in the germinal center B-cell subtype of diffuse large B-cell lymphoma, a single point mutation in the histone methyl-transferase *EZH2* is recurrently mutated in 21.7% patients [328]. The DNA methyl transferase *DNMT3A* is mutated in 22.1% of acute myeloid leukemias [329]. Components of the SWI/SNF nucleosome remodeling complex are mutated in ~19% of cancers [330] such as *hSNF5/INI1* in rhabdoid tumors, ARID1A in colon cancer [315], and ARID2 in hepatocellular carcinoma [316]. This class of epigenetic modifier mutations implicates a causative biological role of epigenetic information in oncogenesis. In fact, a direct analogy can be made between how DNA-repair mutations accelerate the acquisition of additional mutations in cancer, and epigenetic regulator mutations accelerate epimutation in cancer. It therefore stands to reason that epigenetic information itself is fundamentally involved in oncogenesis. This is supported in epigenetic studies which show dysregulation exists in the absence of obvious epigenetic modifier mutations [331].

One consequence of global epigenetic dysregulation, is a change to the transcriptional capacity of TEs. As global TE repression is perturbed, a new transcriptional landscape becomes available to the cancer cell. As discussed in Chapter 1, a well characterized example of cancer-specific TE transcription with oncogenic effects is the HL oncogene colony stimulating factor 1 receptor (*CSF1R*), which is natively restricted to the myeloid lineage but this restriction is subverted in HL through the transcriptional activation of a normally dormant ERV LTR as an alternative promoter [150]. *CSF1R* overexpression in HL is oncogenic and correlates with poor patient outcome [152]. The 'exaptation' of an LTR promoter provides the means with which an otherwise fate-restricted proto-oncogene may be accessed.

5.1 Models of onco-exaptation

The aforementioned cases of onco-exaptation discussed in this thesis are a distinct mechanism by which proto-oncogenes become oncogenic. Classical activating mutations within TEs may also lead to transcription of downstream oncogenes but we are unaware of any evidence for DNA mutations resulting in LTR/TE transcriptional activation, including cases where local DNA was sequenced [161] and unpublished results]. Thus, it is important to consider the etiology through which LTRs/TEs become incorporated into new regulatory units in cancer. This mechanism could be therapeutically or diagnostically important and perhaps even model how TEs, mainly LTRs, influence genome regulation in evolutionary time.

From the known onco-exaptation cases [205] there is often no or very little detectable transcription from the LTR/TE in any cell type other than the cancer type in which it was reported, suggesting the activity is specific to a particular TE in a particular cancer. In other cases, CAGE or EST data show that the LTR/TE can be expressed in other normal or cancer cell types, perhaps to a lower degree. Hence the term “cancer-specific” should be considered a relative one. Indeed, the

idea that the same TE-promoted gene transcripts occur recurrently in tumors from different individuals is central to understanding how these transcripts arise. Below I present two models that may explain the phenomenon of onco-exaptation.

5.1.1 The de-repression model

Lamprecht and co-workers proposed a ‘De-repression model’ for the LTR driven transcription of *CSF1R* [150]. The distinguishing feature of this model is that onco-exaptations arise deterministically, as a consequence of molecular changes that occur during oncogenesis, changes which act to de-repress LTRs or other TEs (Figure 5.1). It follows that ‘activation’ of normally dormant TEs/LTRs could lead to robust oncogene expression. In the *CSF1R* case, the THE1B LTR, which promotes *CSF1R* in cHL, contains binding sites for the transcription factors Sp1, AP-1 and NF-kB, each of which contributes to promoter activity in a luciferase reporter experiment [150]. High NF-kB activity, which is known to be up-regulated in HL, loss of the epigenetic corepressor CBFA2T3 as well as LTR hypomethylation all correlated with CSF1R-positive HL driven by the LTR [150]. Under the de-repression model, the THE1B LTR is repressed by default in the cell but under a particular set of conditions (gain of NF-kB, loss of CBFA2T3, loss of DNA methylation) the LTR promoter is remodeled into an active state [293]. More generally, the model proposes that a particular LTR activation is a consequence of the pathogenic or disrupted molecular state of the cancer cell. In a similar vein, Weber et al. proposed that the L1-driven transcription of *MET* arose as a consequence of global DNA hypomethylation and loss of repression of TEs in cancer [158].

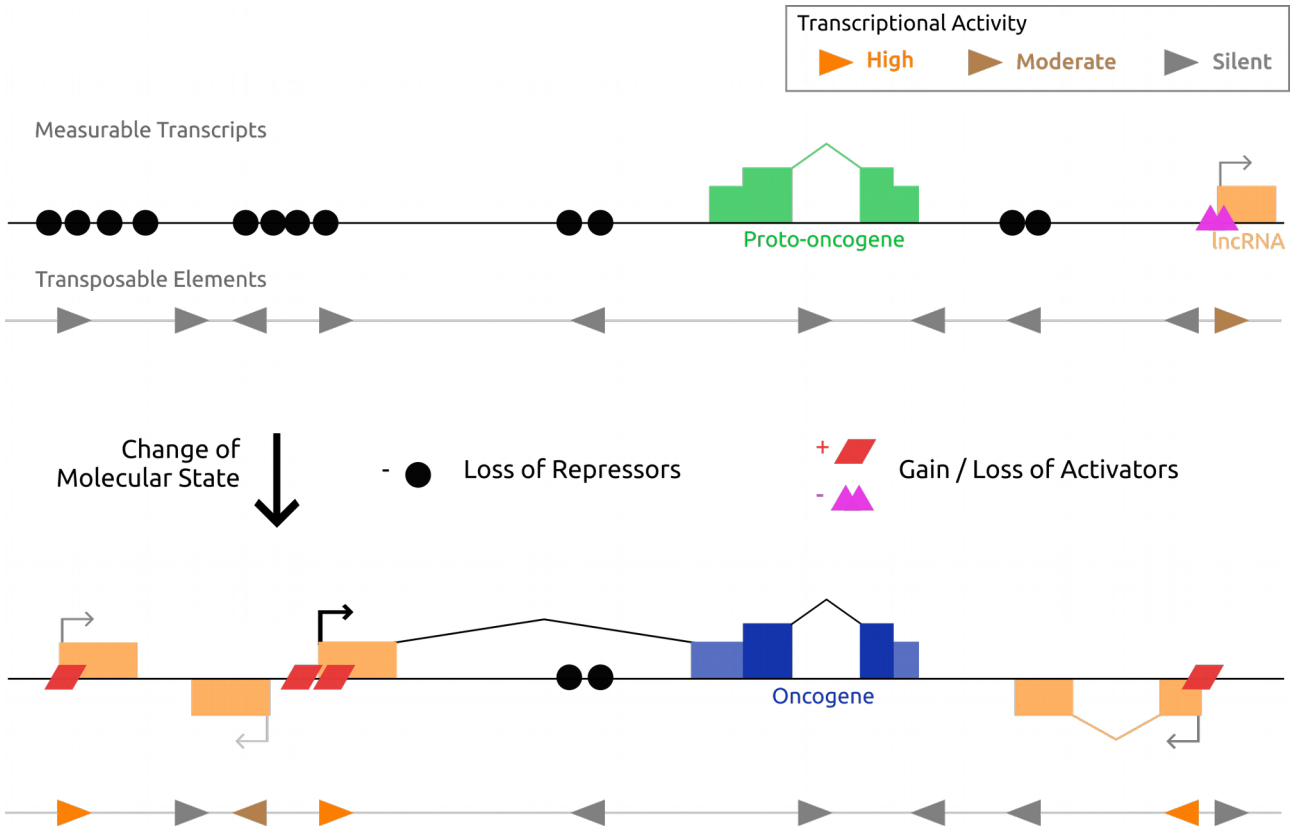


Figure 5.1: De-repression model for onco-exaptation

In the normal or pre-malignant state TEs (grey triangles) are largely silenced across the genome. There is low transcriptional activity to produce long non-coding RNA (orange box), or express coding genes in the case of evolutionary exaptations (not shown). The example proto-oncogene (green box) is under the regulatory control of its native, restrictive promoter. During the process of transformation and/or oncogenesis, a change in the molecular state of the cell occurs leading to loss of TE repressors (black circles), i.e. DNA hypomethylation, loss of transcriptional or epigenetic repressive factors. The change could also be accompanied by a change/gain in activating factor activities (red and purple shapes). Together these de-repression events result in higher TE promoter activity (orange triangles) and more TE-derived transcripts based on the factors that become deregulated. Oncogenic activation of proto-oncogenes is a consequence of a particular molecular milieu that arises in the cancerous cells.

The *LOR1a-IRF5* onco-exaptation in HL (Chapter 4 and [224]) can be interpreted using a de-repression model. An interferon regulatory factor binding element site was created at the intersection of the *LOR1a* LTR and genomic DNA. In normal and cHL cells negative for *LOR1a-*

IRF5, the LTR is methylated and protected from DNAse digestion, a state that is lost in de-repressed cHL cells. This transcription factor-binding motif is responsive to IRF5 itself and creates a positive feedback loop between IRF5 and the chimeric *LOR1a-IRF5* transcript. Thus epigenetic de-repression of this element may reveal an oncogenic exploitation, resulting in high recurrence of *LOR1a* LTR-driven IRF5 in cHL [224].

A de-repression model explains several experimental observations, such as the necessity for a given set of factors to be present (or absent) for a certain promoter to be active, especially when those factors differ between cell states. Indeed, experiments probing the mechanism of TE/LTR activation have used this line of reasoning, often focusing on DNA methylation [150,158,159,166]. The limitation of these studies is that they fail to determine if a given condition is sufficient for onco-exaptation to arise. For instance, the human genome contains >37,000 *THE1* LTR loci (Table 1.2), and indeed this set of LTRs is more active in some HL cells compared to B-cells as (Figure 4.1B). The critical question is why particular *THE1* LTR loci, such as *THE1B-CSF1R*, are recurrently de-repressed in HL, yet thousands of homologous LTRs are not.

5.1.2 The epigenetic evolution model

A central premise in the TE field states that TEs can be beneficial to a host genome since they increase genetic variation in a population and thus increase the rate at which evolution (by natural selection) occurs [332–335]. The epigenetic evolution model for onco-exaptation (Figure 5.2) draws a parallel to this premise within the context of tumor evolution.

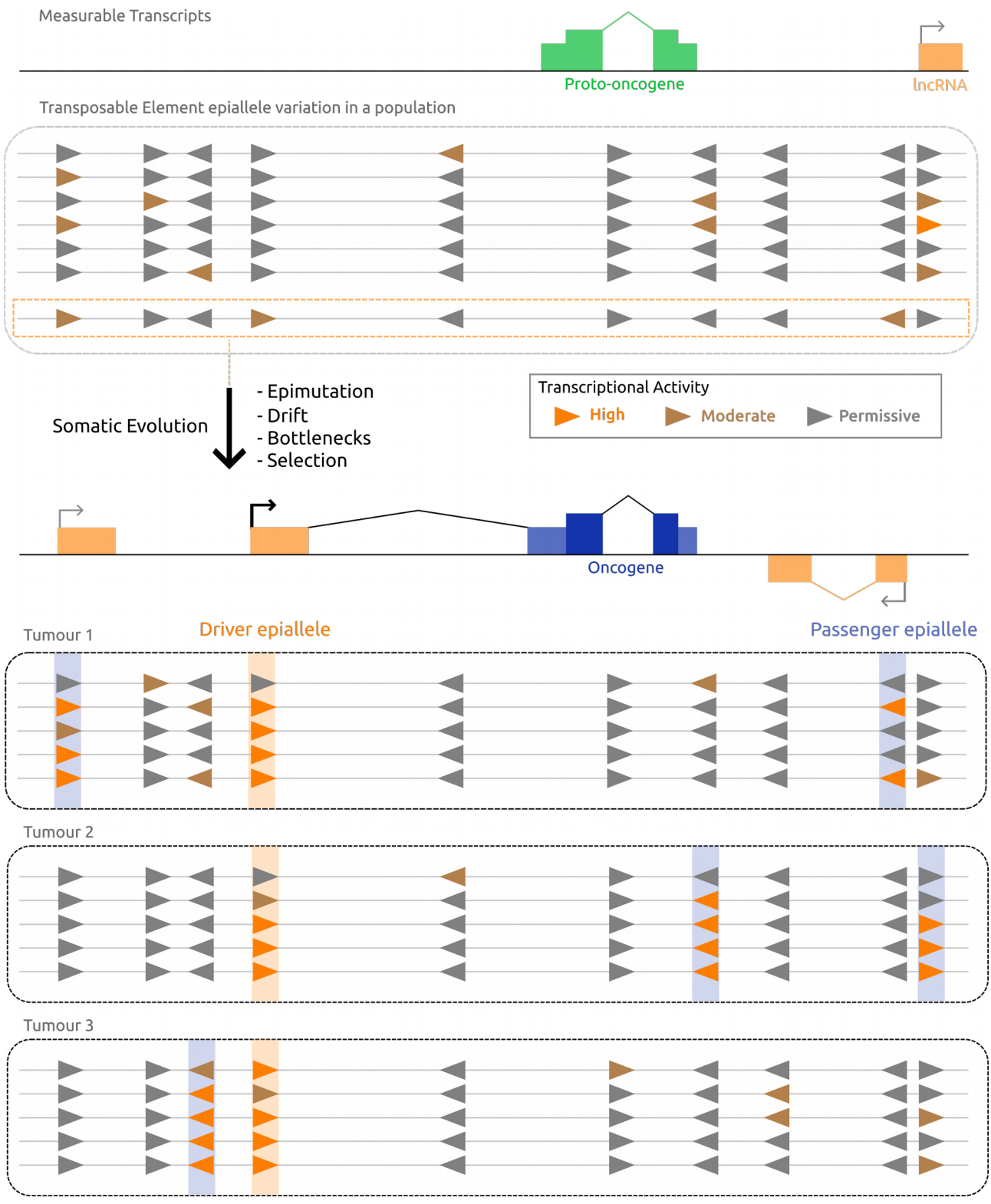


Figure 5.2: Epigenetic evolution model for onco-exaptation

Text 8: Figure 5.2 continued...

In the starting cell population there is a dispersed and low/noisy promoter activity at TEs (colored triangles) from a set of transcriptionally permissive TEs (gray triangles). TE-derived transcript expression is low and variable between cells. Some transcripts are more reliably measurable (orange box). Clonal tumor evolutionary forces change the frequency and expression of TE-derived transcripts by homogenizing epialleles and use of TE promoters (highlighted haplotype). A higher frequency of 'active' TE epialleles at a locus results in increased measurable transcripts initiating from that position. TE epialleles that promote oncogenesis can be selected for and arise multiple times independently as driver epialleles, in contrast to the more dispersed passenger epialleles.

Key to the epigenetic evolution model is that there is high epigenetic variance, both between LTR loci and at the same LTR locus between cells in a population. This epigenetic variance fosters regulatory innovation, and increases during oncogenesis. In accord with this idea are several studies showing that DNA methylation variation, or heterogeneity, increases in tumor cell populations and this isn't simply a global hypomethylation relative to normal cells [336–338](reviewed in [339]). In contrast to the de-repression model, a particular pathogenic molecular state is not sufficient or necessary for TE-driven transcripts to arise; instead the given state only dictates which sets of TEs in the genome are permissive for transcription. Likewise, global de-repression events, such as DNA hypomethylation or mutation of epigenetic regulators, are not necessary, but would increase the rate at which novel transcriptional regulation evolves.

Underpinning this model is the idea that LTRs are highly abundant and self-contained promoters dispersed across the genome that can stochastically initiate low or noisy transcription. This transcriptional noise is a kind of epigenetic variation and thus contributes to cell-cell variation in a population. Indeed, by re-analyzing CAGE data-sets of retrotransposon derived TSSs published by Faulkner *et al*. [75], I observed that TE-derived TSSs have lower expression levels and are less reproducible between biological replicates, compared to non-TE promoters (Figure 2.7). During

malignant transformation, TFs can become deregulated and genome-wide epigenetic perturbations occur [108,340,341] which would change the set of LTRs that are potentially active as well as possibly increasing the total level of LTR-driven transcriptional noise. Up-regulation of specific LTR-driven transcripts would initially be weak and stochastic, from the set of permissive LTRs. Those cells gaining an LTR-driven transcript which confers a growth advantage would then be selected for, and the resultant oncogene expression would increase in the tumor population as that epiallele increases in frequency, in a similar fashion as proposed for the epigenetic silencing of tumor suppressor genes [342–344]. Notably, this scenario also means that within a tumor, LTR-driven transcription would be subject to epigenetic bottleneck effects as well, and that transcriptional LTR noise can become "passenger" expression signals as the cancer cells undergo somatic, clonal expansion.

It may be counter-intuitive to think of evolution and selection as occurring outside the context of genetic variation, but the fact that both genetic mutations and non-genetic/epigenetic variants can contribute to somatic evolution of a cancer is becoming clear [338,345–348]. Epigenetic information or variation by definition is transmitted from mother to daughter cells. Thus, in the specific context of a somatic/asexual cell population such as a tumor, this information, which is both variable between cells in the population and heritable, will be subject to evolutionary changes in frequency. DNA methylation in particular has a well-established mechanism by which information (mainly gene repression) is transmitted epigenetically from mother to daughter cells [349] and DNA hypomethylation at LTRs often correlates with their expression [150,224,268]. Thus, this model suggests that one important type of "epigenetic variant" or epiallele is the transcriptional status of the LTR itself, since the phenotypic impact of LTR transcription may be high in onco-exaptation. Especially in light of the fact that large numbers of these highly

homologous sequences are spread across the genome, epigenetic variation, and possibly selection, at LTRs creates a fascinating system by which epigenetic evolution in cancer may occur.

5.2 Conclusions

In this chapter I have presented two models that may explain onco-exaptation events. These two models are not mutually exclusive but they do provide alternative hypotheses by which TE-driven transcription may be interpreted. This dichotomy is possibly best exemplified by the *ERBB4* case (Figure 1.2E) [129]. There are two LTR-derived promoters which result in aberrant *ERBB4* expression in ALCL. From the de-repression model viewpoint, both LTR elements are grouped MLT1 (MLT1C and MLT1H) and thus this group can be interpreted as derepressed. From the epigenetic evolution model viewpoint, this is convergent evolution/selection for onco-exaptations involving *ERBB4*.

During the final preparation of this thesis, a comprehensive study of TE-initiated transcripts in cancer was published by Jang et al, [350]. This study greatly expands the known repertoire of known TE-initiated transcripts with 625 distinct transcripts between TE and oncogenes. As the list of TE-initiated oncogenes cases grows (and thus the biological relevance of this phenomenon to tumorigenesis), a deeper understanding of the underlying mechanisms and dynamics giving rise to TE-initiated transcription is required.

Through application of the de-repression model,TE-derived transcripts could be used as a diagnostic marker in cancer. If the set of TE/LTR derived transcripts are a deterministic consequence of an underlying oncogenic molecular state, by understanding which set of TEs correspond to which molecular state, it might be possible to assay cancer samples for functional molecular phenotypes. In HL for example, CSF1R status is prognostically important [115] and this is dependent on the transcriptional state of a single *THE1B***.** HL also has been postulated to have a

specific increase in *THE1* LTR transcription genome-wide [294], although this may be in a sub-set of HL (Figure 4.1B). Thus, it's reasonable to hypothesize that the prognostic power can be increased if the transcriptional status of all THE1 LTRs is considered. A set of LTRs can then be interpreted as an *in situ* 'molecular sensor' for aberrant NF-kB function in HL / B-cells for instance.

The epigenetic evolution model proposes that LTR-driven transcripts can be interpreted as a set of epimutations in cancer, similar to how oncogenic mutations are analyzed. Genes that are recurrently (and independently) onco-exapted in multiple different tumors of the same cancer type may be a mark of selective pressure for acquiring that transcript. This is distinct from the more diverse/noisy "passenger LTR" transcription occurring across the genome. These active but "passenger LTRs" may be expressed to a high level within a single tumor population due to epigenetic drift and population bottlenecks but would be more variable across different tumors. Thus, analysis of recurrent and cancer-specific TE-derived transcripts may enrich for genes of significance to tumor biology.

Bibliography

1. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, et al. Initial sequencing and analysis of the human genome. Nature. 2001;409:860–921.

2. Hubley R, Finn RD, Clements J, Eddy SR, Jones TA, Bao W, et al. The Dfam database of repetitive DNA families. Nucleic Acids Res. 2016;44:D81–9.

3. Kidwell MG, Lisch D. Transposable elements as sources of variation in animals and plants. Proc Natl Acad Sci U S A. 1997;94:7704–11.

4. Brookfield JFY. The ecology of the genome - mobile DNA elements and their hosts. Nat Rev Genet. 2005;6:128–36.

5. Mauricio R. Can ecology help genomics: the genome as ecosystem? Genetica. 2005;123:205–9.

6. Burns KH, Boeke JD. Human Transposon Tectonics. Cell. 2012;149:740–52.

7. Gould SJ, Vrba ES. Exaptation—a Missing Term in the Science of Form. Paleobiology. 1982;8:4–15.

8. Brosius J, Gould SJ. On “genomenclature”: a comprehensive (and respectful) taxonomy for pseudogenes and other “junk DNA”. Proc Natl Acad Sci U S A. 1992;89:10706–10.

9. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012;489:57–74.

10. Jacob F. Evolution and tinkering. Science. 1977;196:1161–6.

11. Laubichler MD. Tinkering: a conceptual and historical evaluation. Novartis Found Symp. 2007;284:20-29; discussion 29-34, 110–5.

12. RepeatMasker Home Page [Internet]. [cited 2015 Jul 22]. Available from: http://www.repeatmasker.org/

13. Kojima KK. Human transposable elements in Repbase: genomic footprints from fish to humans. Mob DNA [Internet]. 2018 [cited 2019 Apr 24];9. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5753468/

14. Tarlinton R, Meers J, Young P. Biology and evolution of the endogenous koala retrovirus. Cell Mol Life Sci CMLS. 2008;65:3413–21.

15. Young GR, Eksmond U, Salcedo R, Alexopoulou L, Stoye JP, Kassiotis G. Resurrection of endogenous retroviruses in antibody-deficient mice. Nature. 2012;491:774–8.

16. Belshaw R, Watson J, Katzourakis A, Howe A, Woolven-Allen J, Burt A, et al. Rate of Recombinational Deletion among Human Endogenous Retroviruses. J Virol. 2007;81:9437–42.

17. Gemmell P, Hein J, Katzourakis A. Phylogenetic Analysis Reveals That ERVs “Die Young” but HERV-H Is Unusually Conserved. PLOS Comput Biol. 2016;12:e1004964.

18. Magiorkinis G, Gifford RJ, Katzourakis A, Ranter JD, Belshaw R. Env-less endogenous retroviruses are genomic superspreaders. Proc Natl Acad Sci. 2012;109:7385–90.

19. Ribet D, Harper F, Dewannieux M, Pierron G, Heidmann T. Murine MusD Retrotransposon: Structure and Molecular Evolution of an “Intracellularized” Retrovirus. J Virol. 2007;81:1888–98.

20. Maksakova IA, Romanish MT, Gagnier L, Dunn CA, van de Lagemaat LN, Mager DL. Retroviral elements and their hosts: insertional mutagenesis in the mouse germ line. PLoS Genet. 2006;2:e2.

21. Gagnier L, Belancio VP, Mager DL. Mouse germ line mutations due to retrotransposon insertions. Mob DNA. 2019;10:15.

22. Magiorkinis G, Blanco-Melo D, Belshaw R. The decline of human endogenous retroviruses: extinction and survival. Retrovirology. 2015;12:8.

23. Bannert N, Kurth R. The evolutionary dynamics of human endogenous retroviral families. Annu Rev Genomics Hum Genet. 2006;7:149–73.

24. Chen J-M, Stenson PD, Cooper DN, Férec C. A systematic analysis of LINE-1 endonuclease-dependent retrotranspositional events causing human genetic disease. Hum Genet. 2005;117:411–27.

25. Turner G, Barbulescu M, Su M, Jensen-Seaman MI, Kidd KK, Lenz J. Insertional polymorphisms of full-length endogenous retroviruses in humans. Curr Biol. 2001;11:1531–5.

26. Brouha B, Schustak J, Badge RM, Lutz-Prigge S, Farley AH, Moran JV, et al. Hot L1s account for the bulk of retrotransposition in the human population. Proc Natl Acad Sci U S A. 2003;100:5280–5.

27. Deininger PL, Moran JV, Batzer MA, Kazazian HH. Mobile elements and mammalian genome evolution. Curr Opin Genet Dev. 2003;13:651–8.

28. Denli AM, Narvaiza I, Kerman BE, Pena M, Benner C, Marchetto MCN, et al. Primate-Specific ORF0 Contributes to Retrotransposon-Mediated Diversity. Cell. 2015;163:583–93.

29. Curcio MJ, Derbyshire KM. The outs and ins of transposition: from Mu to Kangaroo. Nat Rev Mol Cell Biol. 2003;4:865–77.

30. Kazazian JH. An estimated frequency of endogenous insertional mutations in humans. Nat Genet. 1999;22:130–130.

31. Coufal NG, Garcia-Perez JL, Peng GE, Yeo GW, Mu Y, Lovci MT, et al. L1 retrotransposition in human neural progenitor cells. Nature. 2009;460:1127–31.

32. Evrony GD, Cai X, Lee E, Hills LB, Elhosary PC, Lehmann HS, et al. Single-neuron sequencing analysis of L1 retrotransposition and somatic mutation in the human brain. Cell. 2012;151:483–96.

33. Baillie JK, Barnett MW, Upton KR, Gerhardt DJ, Richmond TA, De Sapio F, et al. Somatic retrotransposition alters the genetic landscape of the human brain. Nature. 2011;479:534–7.

34. Faulkner GJ, Garcia-Perez JL. L1 Mosaicism in Mammals: Extent, Effects, and Evolution. Trends Genet TIG. 2017;33:802–16.

35. Faulkner GJ, Billon V. L1 retrotransposition in the soma: a field jumping ahead. Mob DNA [Internet]. 2018 [cited 2019 Jan 24];9. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6035798/

36. Kramerov DA, Vassetzky NS. Origin and evolution of SINEs in eukaryotic genomes. Heredity. 2011;107:487–95.

37. Longo MS, Brown JD, Zhang C, O'Neill MJ, O'Neill RJ. Identification of a recently active mammalian SINE derived from ribosomal RNA. Genome Biol Evol. 2015;7:775–88.

38. Kojima KK. A New Class of SINEs with snRNA Gene-Derived Heads. Genome Biol Evol. 2015;7:1702–12.

39. Ullu E, Tschudi C. Alu sequences are processed 7SL RNA genes. Nature. 1984;312:171–2.

40. Bovia F, Wolff N, Ryser S, Strub K. The SRP9/14 subunit of the human signal recognition particle binds to a variety of Alu-like RNAs and with higher affinity than its mouse homolog. Nucleic Acids Res. 1997;25:318–26.

41. Deininger P. Alu elements: know the SINEs. Genome Biol. 2011;12:236.

42. Cordaux R, Hedges DJ, Herke SW, Batzer MA. Estimating the retrotransposition rate of human Alu elements. Gene. 2006;373:134–7.

43. Jurka J, Zietkiewicz E, Labuda D. Ubiquitous mammalian-wide interspersed repeats (MIRs) are molecular fossils from the mesozoic era. Nucleic Acids Res. 1995;23:170–5.

44. Jjingo D, Conley AB, Wang J, Mariño-Ramírez L, Lunyak VV, Jordan IK. Mammalian-wide interspersed repeat (MIR)-derived enhancers and the regulation of human gene expression. Mob DNA. 2014;5:14.

45. Ostertag EM, Goodier JL, Zhang Y, Kazazian HH. SVA elements are nonautonomous retrotransposons that cause disease in humans. Am J Hum Genet. 2003;73:1444–51.

46. Hancks DC, Kazazian HH. Roles for retrotransposon insertions in human disease. Mob DNA. 2016;7:9.

47. Pace JK, Feschotte C. The evolutionary history of human DNA transposons: Evidence for intense activity in the primate lineage. Genome Res. 2007;17:422–32.

48. Ray DA, Pagan HJT, Thompson ML, Stevens RD. Bats with hATs: Evidence for Recent DNA Transposon Activity in Genus Myotis. Mol Biol Evol. 2007;24:632–9.

49. Mitra R, Li X, Kapusta A, Mayhew D, Mitra RD, Feschotte C, et al. Functional characterization of piggyBat from the bat Myotis lucifugus unveils an active mammalian DNA transposon. Proc Natl Acad Sci U S A. 2013;110:234–9.

50. Zhang Y, Romanish MT, Mager DL. Distributions of Transposable Elements Reveal Hazardous Zones in Mammalian Introns. PLOS Comput Biol. 2011;7:e1002046.

51. Zhang Y, Babaian A, Gagnier L, Mager DL. Visualized computational predictions of transcriptional effects by intronic endogenous retroviruses. PloS One. 2013;8:e71971.

52. Boissinot S, Davis J, Entezam A, Petrov D, Furano AV. Fitness cost of LINE-1 (L1) activity in humans. Proc Natl Acad Sci. 2006;103:9590–4.

53. Rasgon JL, Gould F. Transposable element insertion location bias and the dynamics of gene drive in mosquito populations. Insect Mol Biol. 2005;14:493–500.

54. Castillo DM, Mell JC, Box KS, Blumenstiel JP. Molecular evolution under increasing transposable element burden in Drosophila: A speed limit on the evolutionary arms race. BMC Evol Biol. 2011;11:258.

55. Platt RN, Vandewege MW, Ray DA. Mammalian transposable elements and their impacts on genome evolution. Chromosome Res. 2018;26:25–43.

56. Rebollo R, Romanish MT, Mager DL. Transposable elements: an abundant and natural source of regulatory sequences for host genes. Annu Rev Genet. 2012;46:21–42.

57. Jangam D, Feschotte C, Betrán E. Transposable Element Domestication As an Adaptation to Evolutionary Conflicts. Trends Genet. 2017;33:817–31.

58. Carbone L, Alan Harris R, Gnerre S, Veeramah KR, Lorente-Galdos B, Huddleston J, et al. Gibbon genome and the fast karyotype evolution of small apes. Nature. 2014;513:195–201.

59. Ricci M, Peona V, Guichard E, Taccioli C, Boattini A. Transposable Elements Activity is Positively Related to Rate of Speciation in Mammals. J Mol Evol. 2018;86:303–10.

60. Aravin AA, Hannon GJ, Brennecke J. The Piwi-piRNA Pathway Provides an Adaptive Defense in the Transposon Arms Race. Science. 2007;318:761–4.

61. Yang P, Wang Y, Macfarlan TS. The Role of KRAB-ZFPs in Transposable Element Repression and Mammalian Evolution. Trends Genet TIG. 2017;33:871–81.

62. Böhnlein S, Hauber J, Cullen BR. Identification of a U5-specific sequence required for efficient polyadenylation within the human immunodeficiency virus long terminal repeat. J Virol. 1989;63:421–4.

63. Brin E, Leis J. HIV-1 Integrase Interaction with U3 and U5 Terminal Sequences in Vitro Defined Using Substrates with Random Sequences. J Biol Chem. 2002;277:18357–64.

64. Kang SY, Ahn DG, Lee C, Lee YS, Shin C-G. Functional nucleotides of U5 LTR determining substrate specificity of prototype foamy virus integrase. J Microbiol Biotechnol. 2008;18:1044–9.

65. Karn J, Stoltzfus CM. Transcriptional and Posttranscriptional Regulation of HIV-1 Gene Expression. Cold Spring Harb Perspect Med [Internet]. 2012 [cited 2018 Aug 10];2. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3281586/

66. Chuong EB, Elde NC, Feschotte C. Regulatory activities of transposable elements: from conflicts to benefits. Nat Rev Genet. 2017;18:71–86.

67. Kunarso G, Chia N-Y, Jeyakani J, Hwang C, Lu X, Chan Y-S, et al. Transposable elements have rewired the core regulatory network of human embryonic stem cells. Nat Genet. 2010;42:631–4.

68. Cui F, Sirotin MV, Zhurkin VB. Impact of Alu repeats on the evolution of human p53 binding sites. Biol Direct. 2011;6:2.

69. Lev-Maor G, Ram O, Kim E, Sela N, Goren A, Levanon EY, et al. Intronic Alus Influence Alternative Splicing. PLOS Genet. 2008;4:e1000204.

70. Targovnik HM, Pohl V, Christophe D, Cabrer B, Brocas H, Vassart G. Structural organization of the 5’ region of the human thyroglobulin gene. Eur J Biochem. 1984;141:271–7.

71. Sela N, Mersch B, Hotz-Wagenblatt A, Ast G. Characteristics of transposable element exonization within human and mouse. PloS One. 2010;5:e10907.

72. Schmitz J, Brosius J. Exonization of transposed elements: A challenge and opportunity for evolution. Biochimie. 2011;93:1928–34.

73. Deininger P, Morales ME, White TB, Baddoo M, Hedges DJ, Servant G, et al. A comprehensive approach to expression of L1 loci. Nucleic Acids Res. 2017;45:e31.

74. Farré D, Engel P, Angulo A. Novel Role of 3'UTR-Embedded Alu Elements as Facilitators of Processed Pseudogene Genesis and Host Gene Capture by Viral Genomes. PloS One. 2016;11:e0169196.

75. Faulkner GJ, Kimura Y, Daub CO, Wani S, Plessy C, Irvine KM, et al. The regulated retrotransposon transcriptome of mammalian cells. Nat Genet. 2009;41:563–71.

76. Kapusta A, Kronenberg Z, Lynch VJ, Zhuo X, Ramsay L, Bourque G, et al. Transposable Elements Are Major Contributors to the Origin, Diversification, and Regulation of Vertebrate Long Noncoding RNAs. PLoS Genet. 2013;9:e1003470.

77. Kelley D, Rinn J. Transposable elements reveal a stem cell-specific class of long noncoding RNAs. Genome Biol. 2012;13:R107.

78. Trizzino M, Kapusta A, Brown CD. Transposable elements generate regulatory novelty in a tissue-specific fashion. BMC Genomics [Internet]. 2018 [cited 2018 Sep 11];19. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6006921/

79. Walsh CP, Chaillet JR, Bestor TH. Transcription of IAP endogenous retroviruses is constrained by cytosine methylation. Nat Genet. 1998;20:116–7.

80. Szpakowski S, Sun X, Lage JM, Dyer A, Rubinstein J, Kowalski D, et al. Loss of epigenetic silencing in tumors preferentially affects primate-specific retroelements. Gene. 2009;448:151–67.

81. Gerdes P, Richardson SR, Mager DL, Faulkner GJ. Transposable elements in the mammalian embryo: pioneers surviving through stealth and service. Genome Biol. 2016;17:100.

82. Maksakova IA, Thompson PJ, Goyal P, Jones SJ, Singh PB, Karimi MM, et al. Distinct roles of KAP1, HP1 and G9a/GLP in silencing of the two-cell-specific retrotransposon MERVL in mouse ES cells. Epigenetics Chromatin. 2013;6:15.

83. Ecco G, Cassano M, Kauzlaric A, Duc J, Coluccio A, Offner S, et al. Transposable elements and their KRAB-ZFP controllers regulate gene expression in adult tissues. Dev Cell. 2016;36:611–23.

84. Rowe HM, Jakobsson J, Mesnard D, Rougemont J, Reynard S, Aktas T, et al. KAP1 controls endogenous retroviruses in embryonic stem cells. Nature. 2010;463:237–40.

85. Groh S, Schotta G. Silencing of endogenous retroviruses by heterochromatin. Cell Mol Life Sci. 2017;74:2055–65.

86. Koito A, Ikeda T. Intrinsic immunity against retrotransposons by APOBEC cytidine deaminases. Front Microbiol. 2013;4:28.

87. Sienski G, Dönertas D, Brennecke J. Transcriptional silencing of transposons by Piwi and maelstrom and its impact on chromatin state and gene expression. Cell. 2012;151:964–80.

88. Le Thomas A, Rogers AK, Webster A, Marinov GK, Liao SE, Perkins EM, et al. Piwi induces piRNA-guided transcriptional silencing and establishment of a repressive chromatin state. Genes Dev. 2013;27:390–9.

89. Ishizu H, Siomi H, Siomi MC. Biology of PIWI-interacting RNAs: new insights into biogenesis and function inside and outside of germlines. Genes Dev. 2012;26:2361–73.

90. Jacobs FMJ, Greenberg D, Nguyen N, Haeussler M, Ewing AD, Katzman S, et al. An evolutionary arms race between KRAB zinc-finger genes *ZNF91/93* and SVA/L1 retrotransposons. Nature. 2014;516:242–5.

91. Flajnik MF, Kasahara M. Origin and evolution of the adaptive immune system: genetic events and selective pressures. Nat Rev Genet. 2010;11:47–59.

92. Sakano H, Hüppi K, Heinrich G, Tonegawa S. Sequences at the somatic recombination sites of immunoglobulin light-chain genes. Nature. 1979;280:288–94.

93. Agrawal A, Eastman QM, Schatz DG. Transposition mediated by RAG1 and RAG2 and its implications for the evolution of the immune system. Nature. 1998;394:744–51.

94. Huang S, Tao X, Yuan S, Zhang Y, Li P, Beilinson HA, et al. Discovery of an Active RAG Transposon Illuminates the Origins of V(D)J Recombination. Cell. 2016;166:102–14.

95. Dupressoir A, Vernochet C, Bawa O, Harper F, Pierron G, Opolon P, et al. Syncytin-A knockout mice demonstrate the critical role in placentation of a fusogenic, endogenous retrovirus-derived, envelope gene. Proc Natl Acad Sci U S A. 2009;106:12127–32.

96. Dupressoir A, Lavialle C, Heidmann T. From ancestral infectious retroviruses to bona fide cellular genes: role of the captured syncytins in placentation. Placenta. 2012;33:663–71.

97. Cornelis G, Vernochet C, Carradec Q, Souquere S, Mulot B, Catzeflis F, et al. Retroviral envelope gene captures and syncytin exaptation for placentation in marsupials. Proc Natl Acad Sci. 2015;112:E487–96.

98. Cornelis G, Funk M, Vernochet C, Leal F, Tarazona OA, Meurice G, et al. An endogenous retroviral envelope syncytin and its cognate receptor identified in the viviparous placental Mabuya lizard. Proc Natl Acad Sci U S A. 2017;114:E10991–1000.

99. Cohen CJ, Lock WM, Mager DL. Endogenous retroviral LTRs as promoters for human genes: a critical assessment. Gene. 2009;448:105–14.

100. Rebollo R, Farivar S, Mager DL. C-GATE - catalogue of genes affected by transposable elements. Mob DNA. 2012;3:9.

101. Cohen CJ, Rebollo R, Babovic S, Dai EL, Robinson WP, Mager DL. Placenta-specific Expression of the Interleukin-2 (IL-2) Receptor β Subunit from an Endogenous Retroviral Promoter. J Biol Chem. 2011;286:35543–52.

102. Dunn-Fletcher CE, Muglia LM, Pavlicev M, Wolf G, Sun M-A, Hu Y-C, et al. Anthropoid primate–specific retroviral element THE1B controls expression of CRH in placenta and alters gestation length. PLoS Biol [Internet]. 2018 [cited 2019 Jan 24];16. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6166974/

103. Chuong EB, Rumi MAK, Soares MJ, Baker JC. Endogenous retroviruses function as species-specific enhancer elements in the placenta. Nat Genet. 2013;45:325–9.

104. Chuong EB, Elde NC, Feschotte C. Regulatory evolution of innate immunity through co-option of endogenous retroviruses. Science. 2016;351:1083–7.

105. Fuentes DR, Swigut T, Wysocka J. Systematic perturbation of retroviral LTRs reveals widespread long-range effects on human gene regulation. eLife. 2018;7.

106. McCLINTOCK B. The origin and behavior of mutable loci in maize. Proc Natl Acad Sci U S A. 1950;36:344–55.

107. Jones RN. McClintock's controlling elements: the full story. Cytogenet Genome Res. 2005;109:90–103.

108. Hanahan D, Weinberg RA. Hallmarks of cancer: the next generation. Cell. 2011;144:646–74.

109. Kaer K, Speek M. Retroelements in human disease. Gene. 2013;518:231–41.

110. Miki Y, Nishisho I, Horii A, Miyoshi Y, Utsunomiya J, Kinzler KW, et al. Disruption of the APC gene by a retrotransposal insertion of L1 sequence in a colon cancer. Cancer Res. 1992;52:643–5.

111. Lee E, Iskow R, Yang L, Gokcumen O, Haseley P, Luquette LJ, et al. Landscape of somatic retrotransposition in human cancers. Science. 2012;337:967–71.

112. Solyom S, Ewing AD, Rahrmann EP, Doucet T, Nelson HH, Burns MB, et al. Extensive somatic L1 retrotransposition in colorectal tumors. Genome Res. 2012;22:2328–38.

113. Shukla R, Upton KR, Muñoz-Lopez M, Gerhardt DJ, Fisher ME, Nguyen T, et al. Endogenous retrotransposition activates oncogenic pathways in hepatocellular carcinoma. Cell. 2013;153:101–11.

114. Tubio JMC, Li Y, Ju YS, Martincorena I, Cooke SL, Tojo M, et al. Mobile DNA in cancer. Extensive transduction of nonrepetitive DNA mediated by L1 retrotransposition in cancer genomes. Science. 2014;345:1251343.

115. Rodić N, Steranka JP, Makohon-Moore A, Moyer A, Shen P, Sharma R, et al. Retrotransposon insertions in the clonal evolution of pancreatic ductal adenocarcinoma. Nat Med. 2015;21:1060–4.

116. Ewing AD, Gacita A, Wood LD, Ma F, Xing D, Kim M-S, et al. Widespread somatic L1 retrotransposition occurs early during gastrointestinal cancer evolution. Genome Res. 2015;25:1536–45.

117. Scott EC, Gardner EJ, Masood A, Chuang NT, Vertino PM, Devine SE. A hot L1 retrotransposon evades somatic repression and initiates human colorectal cancer. Genome Res. 2016;26:745–55.

118. Helman E, Lawrence MS, Stewart C, Sougnez C, Getz G, Meyerson M. Somatic retrotransposition in human cancer revealed by whole-genome and exome sequencing. Genome Res. 2014;24:1053–63.

119. Wheelan SJ, Aizawa Y, Han JS, Boeke JD. Gene-breaking: A new paradigm for human retrotransposon-mediated gene evolution. Genome Res. 2005;15:1073–8.

120. Narasimhan VM, Rahbari R, Scally A, Wuster A, Mason D, Xue Y, et al. Estimating the human mutation rate from autozygous segments reveals population differences in human mutational processes. Nat Commun [Internet]. 2017 [cited 2018 Sep 26];8. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5566399/

121. Burns KH. Transposable elements in cancer. Nat Rev Cancer. 2017;17:415–24.

122. Vogt PK. Retroviral Oncogenes: A Historical Primer. Nat Rev Cancer. 2012;12:639–48.

123. Rous P. A SARCOMA OF THE FOWL TRANSMISSIBLE BY AN AGENT SEPARABLE FROM THE TUMOR CELLS. J Exp Med. 1911;13:397–411.

124. Weiss RA, Vogt PK. 100 years of Rous sarcoma virus. J Exp Med. 2011;208:2351–5.

125. Stehelin D, Varmus HE, Bishop JM, Vogt PK. DNA related to the transforming gene(s) of avian sarcoma viruses is present in normal avian DNA. Nature. 1976;260:170.

126. Spector DH, Varmus HE, Bishop JM. Nucleotide sequences related to the transforming gene of avian sarcoma virus are present in DNA of uninfected vertebrates. Proc Natl Acad Sci U S A. 1978;75:4102–6.

127. Martin GS. The road to Src. Oncogene. 2004;23:7910–7.

128. Watson DK, Reddy EP, Duesberg PH, Papas TS. Nucleotide sequence analysis of the chicken c-myc gene reveals homologous and unique coding regions by comparison with the transforming gene of avian myelocytomatosis virus MC29, delta gag-myc. Proc Natl Acad Sci U S A. 1983;80:2146–50.

129. Shields A, Rosenberg N, Baltimore D. Virus production by Abelson murine leukemia virus-transformed lymphoid cells. J Virol. 1979;31:557–67.

130. Dale B, Ozanne B. Characterization of mouse cellular deoxyribonucleic acid homologous to Abelson murine leukemia virus-specific sequences. Mol Cell Biol. 1981;1:731–42.

131. Scolnick EM, Maryak JM, Parks WP. Levels of rat cellular RNA homologous to either Kirsten sarcoma virus or rat type-C virus in cell lines derived from Osborne-Mendel rats. J Virol. 1974;14:1435–44.

132. Anderson GR, Robbins KC. Rat sequences of the Kirsten and Harvey murine sarcoma virus genomes: nature, origin, and expression in rat tumor RNA. J Virol. 1976;17:335–51.

133. Staal SP, Hartley JW. Thymic lymphoma induction by the AKT8 murine retrovirus. J Exp Med. 1988;167:1259–64.

134. Manger R, Najita L, Nichols EJ, Hakomori S, Rohrschneider L. Cell surface expression of the McDonough strain of feline sarcoma virus fms gene product (gp 140fms). Cell. 1984;39:327–37.

135. Hayward WS, Neel BG, Astrin SM. Activation of a cellular onc gene by promoter insertion in ALV-induced lymphoid leukosis. Nature. 1981;290:475–80.

136. Nusse R, Varmus HE. Many tumors induced by the mouse mammary tumor virus contain a provirus integrated in the same region of the host genome. Cell. 1982;31:99–109.

137. Theodorou V, Kimm MA, Boer M, Wessels L, Theelen W, Jonkers J, et al. MMTV insertional mutagenesis identifies genes, gene families and pathways involved in mammary cancer. Nat Genet. 2007;39:759–69.

138. Tsichlis PN. Oncogenesis by Moloney murine leukemia virus. Anticancer Res. 1987;7:171–80.

139. Fan H. Leukemogenesis by Moloney murine leukemia virus: a multistep process. Trends Microbiol. 1997;5:74–82.

140. Leslie KB, Lee F, Schrader JW. Intracisternal A-type particle-mediated activations of cytokine genes in a murine myelomonocytic leukemia: generation of functional cytokine mRNAs by retroviral splicing events. Mol Cell Biol. 1991;11:5562–70.

141. Yarchoan R, Uldrick TS. HIV-Associated Cancers and Related Diseases. N Engl J Med. 2018;378:1029–41.

142. Galli UM, Sauter M, Lecher B, Maurer S, Herbst H, Roemer K, et al. Human endogenous retrovirus rec interferes with germ cell development in mice and may cause carcinoma in situ, the predecessor lesion of germ cell tumors. Oncogene. 2005;24:3223–8.

143. Denne M, Sauter M, Armbruester V, Licht JD, Roemer K, Mueller-Lantzsch N. Physical and functional interactions of human endogenous retrovirus proteins Np9 and rec with the promyelocytic leukemia zinc finger protein. J Virol. 2007;81:5607–16.

144. Kaufmann S, Sauter M, Schmitt M, Baumert B, Best B, Boese A, et al. Human endogenous retrovirus protein Rec interacts with the testicular zinc-finger protein and androgen receptor. J Gen Virol. 2010;91:1494–502.

145. Verdonck K, González E, Van Dooren S, Vandamme A-M, Vanham G, Gotuzzo E. Human T-lymphotropic virus 1: recent knowledge about an ancient infection. Lancet Infect Dis. 2007;7:266–81.

146. Ahmadi Ghezeldasht S, Shirdel A, Assarehzadegan MA, Hassannia T, Rahimi H, Miri R, et al. Human T Lymphotropic Virus Type I (HTLV-I) Oncogenesis: Molecular Aspects of Virus and Host Interactions in Pathogenesis of Adult T cell Leukemia/Lymphoma (ATL). Iran J Basic Med Sci. 2013;16:179–95.

147. Hashimoto K, Suzuki AM, Dos Santos A, Desterke C, Collino A, Ghisletti S, et al. CAGE profiling of ncRNAs in hepatocellular carcinoma reveals widespread activation of retroviral LTR promoters in virus-induced tumors. Genome Res. 2015;25:1812–24.

148. Gibb EA, Warren RL, Wilson GW, Brown SD, Robertson GA, Morin GB, et al. Activation of an endogenous retrovirus-associated long non-coding RNA in human adenocarcinoma. Genome Med. 2015;7:22.

149. Beyer U, Krönung SK, Leha A, Walter L, Dobbelstein M. Comprehensive identification of genes driven by ERV9-LTRs reveals TNFRSF10B as a re-activatable mediator of testicular cancer cell death. Cell Death Differ. 2016;23:64–75.

150. Lamprecht B, Walter K, Kreher S, Kumar R, Hummel M, Lenze D, et al. Derepression of an endogenous long terminal repeat activates the CSF1R proto-oncogene in human lymphoma. Nat Med. 2010;16:571–579, 1p following 579.

151. Jurka J, Kapitonov VV, Pavlicek A, Klonowski P, Kohany O, Walichiewicz J. Repbase Update, a database of eukaryotic repetitive elements. Cytogenet Genome Res. 2005;110:462–7.

152. Steidl C, Diepstra A, Lee T, Chan FC, Farinha P, Tan K, et al. Gene expression profiling of microdissected Hodgkin Reed-Sternberg cells correlates with treatment outcome in classical Hodgkin lymphoma. Blood. 2012;120:3530–40.

153. Martín-Moreno AM, Roncador G, Maestre L, Mata E, Jiménez S, Martínez-Torrecuadrada JL, et al. CSF1R Protein Expression in Reactive Lymphoid Tissues and Lymphoma: Its Relevance in Classical Hodgkin Lymphoma. PloS One. 2015;10:e0125203.

154. Introna M, Luchetti M, Castellano M, Arsura M, Golay J. The myb oncogene family of transcription factors: potent regulators of hematopoietic cell proliferation and differentiation. Semin Cancer Biol. 1994;5:113–24.

155. Ramsay RG, Gonda TJ. MYB function in normal and cancer cells. Nat Rev Cancer. 2008;8:523–34.

156. Mätlik K, Redik K, Speek M. L1 antisense promoter drives tissue-specific transcription of human genes. J Biomed Biotechnol. 2006;2006:71753.

157. Wolff EM, Byun H-M, Han HF, Sharma S, Nichols PW, Siegmund KD, et al. Hypomethylation of a LINE-1 promoter activates an alternate transcript of the MET oncogene in bladders with cancer. PLoS Genet. 2010;6:e1000917.

158. Weber B, Kimhi S, Howard G, Eden A, Lyko F. Demethylation of a LINE-1 antisense promoter in the cMet locus impairs Met signalling through induction of illegitimate transcription. Oncogene. 2010;29:5775–84.

159. Hur K, Cejas P, Feliu J, Moreno-Rubio J, Burgos E, Boland CR, et al. Hypomethylation of long interspersed nuclear element-1 (LINE-1) leads to activation of proto-oncogenes in human colorectal cancer metastasis. Gut. 2014;63:635–46.

160. Gao H, Guan M, Sun Z, Bai C. High c-Met expression is a negative prognostic marker for colorectal cancer: a meta-analysis. Tumour Biol J Int Soc Oncodevelopmental Biol Med. 2015;36:515–20.

161. Wiesner T, Lee W, Obenauf AC, Ran L, Murali R, Zhang QF, et al. Alternative transcription initiation leads to expression of a novel ALK isoform in cancer. Nature. 2015;526:453–7.

162. Alshareef A, Irwin MS, Gupta N, Zhang H-F, Haque M, Findlay SD, et al. The absence of a novel intron 19-retaining ALK transcript (ALK-I19) and MYCN amplification correlates with an excellent clinical outcome in neuroblastoma patients. Oncotarget. 2018;9:10698–713.

163. Mariño-Enríquez A, Dal Cin P. ALK as a paradigm of oncogenic promiscuity: different mechanisms of activation and different fusion partners drive tumors of different lineages. Cancer Genet. 2013;206:357–73.

164. Bao W, Kojima KK, Kohany O. Repbase Update, a database of repetitive elements in eukaryotic genomes. Mob DNA. 2015;6:11.

165. FANTOM Consortium and the RIKEN PMI and CLST (DGT), Forrest ARR, Kawaji H, Rehli M, Baillie JK, de Hoon MJL, et al. A promoter-level mammalian expression atlas. Nature. 2014;507:462–70.

166. Scarfò I, Pellegrino E, Mereu E, Kwee I, Agnelli L, Bergaggio E, et al. Identification of a new subclass of ALK-negative ALCL expressing aberrant levels of ERBB4 transcripts. Blood. 2016;127:221–32.

167. Arteaga CL, Engelman JA. ERBB receptors: from oncogene discovery to basic science to mechanism-based cancer therapeutics. Cancer Cell. 2014;25:282–303.

168. Lock FE, Babaian A, Zhang Y, Gagnier L, Kuah S, Weberling A, et al. A novel isoform of IL-33 revealed by screening for transposable element promoted genes in human colorectal cancer. PLOS ONE. 2017;12:e0180659.

169. Li A, Herbst RH, Canner D, Schenkel JM, Smith OC, Kim JY, et al. Longitudinal single cell profiling of regulatory T cells identifies IL-33 as a driver of tumor immunosuppression. bioRxiv. 2019;512905.

170. Obaidat A, Roth M, Hagenbuch B. The expression and function of organic anion transporting polypeptides in normal tissues and in cancer. Annu Rev Pharmacol Toxicol. 2012;52:135–51.

171. Lee W, Belkhiri A, Lockhart AC, Merchant N, Glaeser H, Harris EI, et al. Overexpression of OATP1B3 confers apoptotic resistance in colon cancer. Cancer Res. 2008;68:10315–23.

172. Nagai M, Furihata T, Matsumoto S, Ishii S, Motohashi S, Yoshino I, et al. Identification of a new organic anion transporting polypeptide 1B3 mRNA isoform primarily expressed in human cancerous tissues and cells. Biochem Biophys Res Commun. 2012;418:818–23.

173. Imai S, Kikuchi R, Tsuruya Y, Naoi S, Nishida S, Kusuhara H, et al. Epigenetic regulation of organic anion transporting polypeptide 1B3 in cancer cell lines. Pharm Res. 2013;30:2880–90.

174. Robbez-Masson L, Rowe HM. Retrotransposons shape species-specific embryonic stem cell gene expression. Retrovirology. 2015;12:45.

175. Izsvák Z, Wang J, Singh M, Mager DL, Hurst LD. Pluripotency and the endogenous retrovirus HERVH: Conflict or serendipity? BioEssays News Rev Mol Cell Dev Biol. 2016;38:109–17.

176. Pérot P, Mullins CS, Naville M, Bressan C, Hühns M, Gock M, et al. Expression of young HERV-H loci in the course of colorectal carcinoma and correlation with molecular subtypes. Oncotarget. 2015;6:40095–111.

177. Liang Q, Xu Z, Xu R, Wu L, Zheng S. Expression patterns of non-coding spliced transcripts from human endogenous retrovirus HERV-H elements in colon cancer. PloS One. 2012;7:e29950.

178. Teft WA, Welch S, Lenehan J, Parfitt J, Choi Y-H, Winquist E, et al. OATP1B1 and tumour OATP1B3 modulate exposure, toxicity, and survival after irinotecan-based chemotherapy. Br J Cancer. 2015;112:857–65.

179. Morin RD, Mendez-Lago M, Mungall AJ, Goya R, Mungall KL, Corbett R, et al. Frequent mutation of histone modifying genes in non-Hodgkin lymphoma. Nature. 2011;476:298–303.

180. Lock FE, Rebollo R, Miceli-Royer K, Gagnier L, Kuah S, Babaian A, et al. Distinct isoform of FABP7 revealed by screening for retroelement-activated genes in diffuse large B-cell lymphoma. Proc Natl Acad Sci U S A. 2014;111:E3534-3543.

181. Thumser AE, Moore JB, Plant NJ. Fatty acid binding proteins: tissue-specific functions in health and disease. Curr Opin Clin Nutr Metab Care. 2014;17:124–9.

182. Liu R-Z, Graham K, Glubrecht DD, Lai R, Mackey JR, Godbout R. A fatty acid-binding protein 7/RXRβ pathway enhances survival and proliferation in triple-negative breast cancer. J Pathol. 2012;228:310–21.

183. Guo W, Hu Z, Bao Y, Li Y, Li S, Zheng Q, et al. A LIN28B Tumor-Specific Transcript in Cancer. Cell Rep. 2018;22:2016–25.

184. St Laurent G, Shtokalo D, Dong B, Tackett MR, Fan X, Lazorthes S, et al. VlincRNAs controlled by retroviral elements are a hallmark of pluripotency and cancer. Genome Biol. 2013;14:R73.

185. Prensner JR, Iyer MK, Sahu A, Asangani IA, Cao Q, Patel L, et al. The long noncoding RNA SChLAP1 promotes aggressive prostate cancer and antagonizes the SWI/SNF complex. Nat Genet. 2013;45:1392–8.

186. Masliah-Planchon J, Bièche I, Guinebretière J-M, Bourdeaut F, Delattre O. SWI/SNF chromatin remodeling and human malignancies. Annu Rev Pathol. 2015;10:145–71.

187. Loewer S, Cabili MN, Guttman M, Loh Y-H, Thomas K, Park IH, et al. Large intergenic non-coding RNA-RoR modulates reprogramming of human induced pluripotent stem cells. Nat Genet. 2010;42:1113–7.

188. Wang Y, Xu Z, Jiang J, Xu C, Kang J, Xiao L, et al. Endogenous miRNA sponge lincRNA-RoR regulates Oct4, Nanog, and Sox2 in human embryonic stem cell self-renewal. Dev Cell. 2013;25:69–80.

189. Eades G, Wolfson B, Zhang Y, Li Q, Yao Y, Zhou Q. lincRNA-RoR and miR-145 regulate invasion in triple-negative breast cancer via targeting ARF6. Mol Cancer Res MCR. 2015;13:330–8.

190. Gao S, Wang P, Hua Y, Xi H, Meng Z, Liu T, et al. ROR functions as a ceRNA to regulate Nanog expression by sponging miR-145 and predicts poor prognosis in pancreatic cancer. Oncotarget. 2016;7:1608–18.

191. Zhou P, Sun L, Liu D, Liu C, Sun L. Long Non-Coding RNA lincRNA-ROR Promotes the Progression of Colon Cancer and Holds Prognostic Value by Associating with miR-145. Pathol Oncol Res POR. 2016;22:733–40.

192. Fan J, Xing Y, Wen X, Jia R, Ni H, He J, et al. Long non-coding RNA ROR decoys gene-specific histone methylation to promote tumorigenesis. Genome Biol. 2015;16:139.

193. Huang J, Zhang A, Ho T-T, Zhang Z, Zhou N, Ding X, et al. Linc-RoR promotes c-Myc expression through hnRNP I and AUF1. Nucleic Acids Res. 2016;44:3059–69.

194. Rangel LBA, Sherman-Baust CA, Wernyj RP, Schwartz DR, Cho KR, Morin PJ. Characterization of novel human ovarian cancer-specific transcripts (HOSTs) identified by serial analysis of gene expression. Oncogene. 2003;22:7225–32.

195. Adams BD, Kasinski AL, Slack FJ. Aberrant regulation and function of microRNAs in cancer. Curr Biol CB. 2014;24:R762-776.

196. Gao Y, Meng H, Liu S, Hu J, Zhang Y, Jiao T, et al. LncRNA-HOST2 regulates cell biological behaviors in epithelial ovarian cancer through a mechanism involving microRNA let-7b. Hum Mol Genet. 2015;24:841–52.

197. Yang F, Lyu S, Dong S, Liu Y, Zhang X, Wang O. Expression profile analysis of long noncoding RNA in HER-2-enriched subtype breast cancer by next-generation sequencing and bioinformatics. OncoTargets Ther. 2016;9:761–72.

198. Zeng Z, Bo H, Gong Z, Lian Y, Li X, Li X, et al. AFAP1-AS1, a long noncoding RNA upregulated in lung cancer and promotes invasion and metastasis. Tumour Biol J Int Soc Oncodevelopmental Biol Med. 2016;37:729–37.

199. Deng J, Liang Y, Liu C, He S, Wang S. The up-regulation of long non-coding RNA AFAP1-AS1 is associated with the poor prognosis of NSCLC patients. Biomed Pharmacother Biomedecine Pharmacother. 2015;75:8–11.

200. Wu W, Bhagat TD, Yang X, Song JH, Cheng Y, Agarwal R, et al. Hypomethylation of Noncoding DNA Regions and Overexpression of the Long Noncoding RNA, AFAP1-AS1, in Barrett's Esophagus and Esophageal Adenocarcinoma. Gastroenterology. 2013;144:956–966.e4.

201. Zhang J-Y, Weng M-Z, Song F-B, Xu Y-G, Liu Q, Wu J-Y, et al. Long noncoding RNA AFAP1-AS1 indicates a poor prognosis of hepatocellular carcinoma and promotes cell proliferation and invasion via upregulation of the RhoA/Rac2 signaling. Int J Oncol. 2016;48:1590–8.

202. Guil S, Esteller M. Cis-acting noncoding RNAs: friends and foes. Nat Struct Mol Biol. 2012;19:1068–75.

203. Leucci E, Vendramin R, Spinazzi M, Laurette P, Fiers M, Wouters J, et al. Melanoma addiction to the long non-coding RNA SAMMSON. Nature. 2016;531:518–22.

204. Harris ML, Baxter LL, Loftus SK, Pavan WJ. Sox proteins in melanocyte development and melanoma. Pigment Cell Melanoma Res. 2010;23:496–513.

205. Babaian A, Mager DL. Endogenous retroviral promoter exaptation in human cancer. Mob DNA. 2016;7:24.

206. Panzitt K, Tschernatsch MMO, Guelly C, Moustafa T, Stradner M, Strohmaier HM, et al. Characterization of HULC, a novel gene with striking up-regulation in hepatocellular carcinoma, as noncoding RNA. Gastroenterology. 2007;132:330–42.

207. Li C, Chen J, Zhang K, Feng B, Wang R, Chen L. Progress and Prospects of Long Noncoding RNAs (lncRNAs) in Hepatocellular Carcinoma. Cell Physiol Biochem Int J Exp Cell Physiol Biochem Pharmacol. 2015;36:423–34.

208. Wang X-S, Zhang Z, Wang H-C, Cai J-L, Xu Q-W, Li M-Q, et al. Rapid identification of UCA1 as a very sensitive and specific unique marker for human bladder carcinoma. Clin Cancer Res Off J Am Assoc Cancer Res. 2006;12:4851–8.

209. Wang F, Li X, Xie X, Zhao L, Chen W. UCA1, a non-protein-coding RNA up-regulated in bladder carcinoma and embryo, influencing cell growth and promoting invasion. FEBS Lett. 2008;582:1919–27.

210. Hu J-J, Song W, Zhang S-D, Shen X-H, Qiu X-M, Wu H-Z, et al. HBx-upregulated lncRNA UCA1 promotes cell growth and tumorigenesis by recruiting EZH2 and repressing p27Kip1/CDK2 signaling. Sci Rep. 2016;6:23521.

211. Flockhart RJ, Webster DE, Qu K, Mascarenhas N, Kovalski J, Kretz M, et al. BRAFV600E remodels the melanocyte transcriptome and induces BANCR to regulate melanoma cell migration. Genome Res. 2012;22:1006–14.

212. Guo Q, Zhao Y, Chen J, Hu J, Wang S, Zhang D, et al. BRAF-activated long non-coding RNA contributes to colorectal cancer migration by inducing epithelial-mesenchymal transition. Oncol Lett. 2014;8:869–75.

213. Wang Y, Guo Q, Zhao Y, Chen J, Wang S, Hu J, et al. BRAF-activated long non-coding RNA contributes to cell proliferation and activates autophagy in papillary thyroid carcinoma. Oncol Lett. 2014;8:1947–52.

214. Cruickshanks HA, Vafadar-Isfahani N, Dunican DS, Lee A, Sproul D, Lund JN, et al. Expression of a large LINE-1-driven antisense RNA is linked to epigenetic silencing of the metastasis suppressor gene TFPI-2 in cancer. Nucleic Acids Res. 2013;41:6857–69.

215. Cruickshanks HA, Tufarelli C. Isolation of cancer-specific chimeric transcripts induced by hypomethylation of the LINE-1 antisense promoter. Genomics. 2009;94:397–406.

216. Lo Nigro C, Wang H, McHugh A, Lattanzio L, Matin R, Harwood C, et al. Methylated tissue factor pathway inhibitor 2 (TFPI2) DNA in serum is a biomarker of metastatic melanoma. J Invest Dermatol. 2013;133:1278–85.

217. Cancer Genome Atlas Research Network, Weinstein JN, Collisson EA, Mills GB, Shaw KRM, Ozenberger BA, et al. The Cancer Genome Atlas Pan-Cancer analysis project. Nat Genet. 2013;45:1113–20.

218. Gosenca D, Gabriel U, Steidler A, Mayer J, Diem O, Erben P, et al. HERV-E-mediated modulation of PLA2G4A transcription in urothelial carcinoma. PloS One. 2012;7:e49341.

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

220. Macfarlan TS, Gifford WD, Driscoll S, Lettieri K, Rowe HM, Bonanomi D, et al. Embryonic stem cell potency fluctuates with endogenous retrovirus activity. Nature. 2012;487:57–63.

221. Lu X, Sachs F, Ramsay L, Jacques P-É, Göke J, Bourque G, et al. The retrovirus HERVH is a long noncoding RNA required for human embryonic stem cell identity. Nat Struct Mol Biol. 2014;21:423–5.

222. Wang J, Xie G, Singh M, Ghanbarian AT, Raskó T, Szvetnik A, et al. Primate-specific endogenous retrovirus-driven transcription defines naive-like stem cells. Nature. 2014;516:405–9.

223. Ramsay L, Marchetto MC, Caron M, Chen S-H, Busche S, Kwan T, et al. Conserved expression of transposon-derived non-coding transcripts in primate stem cells. BMC Genomics. 2017;18:214.

224. Babaian A, Romanish MT, Gagnier L, Kuo LY, Karimi MM, Steidl C, et al. Onco-exaptation of an endogenous retroviral LTR drives IRF5 expression in Hodgkin lymphoma. Oncogene. 2016;35:2542–6.

225. Jurka J, Bao W, Kojima KK. Families of transposable elements, population structure and the origin of species. Biol Direct. 2011;6:44.

226. van de Lagemaat LN, Landry J-R, Mager DL, Medstrand P. Transposable elements in mammals promote regulatory variation and diversification of genes with specialized functions. Trends Genet TIG. 2003;19:530–6.

227. Nigumann P, Redik K, Mätlik K, Speek M. Many human genes are transcribed from the antisense promoter of L1 retrotransposon. Genomics. 2002;79:628–34.

228. Huda A, Bushel PR. Widespread Exonization of Transposable Elements in Human Coding Sequences is Associated with Epigenetic Regulation of Transcription. Transcr Open Access. 2013;1.

229. Conley AB, Piriyapongsa J, Jordan IK. Retroviral promoters in the human genome. Bioinforma Oxf Engl. 2008;24:1563–7.

230. Karimi MM, Goyal P, Maksakova IA, Bilenky M, Leung D, Tang JX, et al. DNA methylation and SETDB1/H3K9me3 regulate predominantly distinct sets of genes, retroelements, and chimeric transcripts in mESCs. Cell Stem Cell. 2011;8:676–87.

231. Wang T, Santos JH, Feng J, Fargo DC, Shen L, Riadi G, et al. A Novel Analytical Strategy to Identify Fusion Transcripts between Repetitive Elements and Protein Coding-Exons Using RNA-Seq. PLOS ONE. 2016;11:e0159028.

232. Sokol M, Jessen KM, Pedersen FS. Human endogenous retroviruses sustain complex and cooperative regulation of gene-containing loci and unannotated megabase-sized regions. Retrovirology. 2015;12:32.

233. Pavlicev M, Hiratsuka K, Swaggart KA, Dunn C, Muglia L. Detecting endogenous retrovirus-driven tissue-specific gene transcription. Genome Biol Evol. 2015;7:1082–97.

234. Pérot P, Mugnier N, Montgiraud C, Gimenez J, Jaillard M, Bonnaud B, et al. Microarray-based sketches of the HERV transcriptome landscape. PloS One. 2012;7:e40194.

235. Jin Y, Tam OH, Paniagua E, Hammell M. TEtranscripts: a package for including transposable elements in differential expression analysis of RNA-seq datasets. Bioinforma Oxf Engl. 2015;31:3593–9.

236. Goerner-Potvin P, Bourque G. Computational tools to unmask transposable elements. Nat Rev Genet. 2018;1.

237. Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol. 2013;14:R36.

238. Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, et al. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nat Protoc. 2012;7:562–78.

239. Griebel T, Zacher B, Ribeca P, Raineri E, Lacroix V, Guigó R, et al. Modelling and simulating generic RNA-Seq experiments with the flux simulator. Nucleic Acids Res. 2012;40:10073–83.

240. Steidl C, Shah SP, Woolcock BW, Rui L, Kawahara M, Farinha P, et al. MHC class II transactivator CIITA is a recurrent gene fusion partner in lymphoid cancers. Nature. 2011;471:377–81.

241. Liu Y, Abdul Razak FR, Terpstra M, Chan FC, Saber A, Nijland M, et al. The mutational landscape of Hodgkin lymphoma cell lines determined by whole-exome sequencing. Leukemia. 2014;28:2248–51.

242. Twa DDW, Chan FC, Ben-Neriah S, Woolcock BW, Mottok A, Tan KL, et al. Genomic rearrangements involving programmed death ligands are recurrent in primary mediastinal large B-cell lymphoma. Blood. 2014;123:2062–5.

243. Günther F, Fritsch S. neuralnet: Training of Neural Networks. R J. 2010;2:30–8.

244. Riedmiller M, Braun H. RPROP - A Fast Adaptive Learning Algorithm. Proc. of ISCIS VII), Universitat; 1992.

245. Igel C, Hüsken M. Improving the Rprop Learning Algorithm. 2000.

246. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012;489:57–74.

247. Hillier LW, Reinke V, Green P, Hirst M, Marra MA, Waterston RH. Massively parallel sequencing of the polyadenylated transcriptome of C. elegans. Genome Res. 2009;19:657–66.

248. Staiger D, Simpson GG. Enter exitrons. Genome Biol. 2015;16:136.

249. Oliver KR, Greene WK. Mobile DNA and the TE-Thrust hypothesis: supporting evidence from the primates. Mob DNA. 2011;2:8.

250. Rosenberg N, Jolicoeur P. Retroviral Pathogenesis. In: Coffin JM, Hughes SH, Varmus HE, editors. Retroviruses [Internet]. Cold Spring Harbor (NY): Cold Spring Harbor Laboratory Press; 1997 [cited 2019 Jan 18]. Available from: http://www.ncbi.nlm.nih.gov/books/NBK19378/

251. De Smet C, Loriot A. DNA hypomethylation in cancer: epigenetic scars of a neoplastic journey. Epigenetics Off J DNA Methylation Soc. 2010;5:206–13.

252. Baylin SB, Jones PA. A decade of exploring the cancer epigenome — biological and translational implications. Nat Rev Cancer. 2011;11:726–34.

253. Ross JP, Rand KN, Molloy PL. Hypomethylation of repeated DNA sequences in cancer. Epigenomics. 2010;2:245–69.

254. Romanish MT, Cohen CJ, Mager DL. Potential mechanisms of endogenous retroviral-mediated genomic instability in human cancer. Semin Cancer Biol. 2010;20:246–53.

255. Purcell M, Kruger A, Tainsky MA. Gene expression profiling of replicative and induced senescence. Cell Cycle. 2014;13:3927–37.

256. Seshagiri S, Stawiski EW, Durinck S, Modrusan Z, Storm EE, Conboy CB, et al. Recurrent R-spondin fusions in colon cancer. Nature. 2012;488:660–4.

257. Hollstein M, Sidransky D, Vogelstein B, Harris CC. p53 mutations in human cancers. Science. 1991;253:49–53.

258. Bischoff FZ, Strong LC, Yim SO, Pratt DR, Siciliano MJ, Giovanella BC, et al. Tumorigenic transformation of spontaneously immortalized fibroblasts from patients with a familial cancer syndrome. Oncogene. 1991;6:183–6.

259. Brocks D, Schmidt CR, Daskalakis M, Jang HS, Shah NM, Li D, et al. DNMT and HDAC inhibitors induce cryptic transcription start sites encoded in long terminal repeats. Nat Genet. 2017;49:1052–60.

260. Costas J, Naveira H. Evolutionary history of the human endogenous retrovirus family ERV9. Mol Biol Evol. 2000;17:320–30.

261. Di Cristofano A, Strazzullo M, Longo L, La Mantia G. 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. 1995;23:2823–30.

262. Chen H-J, Carr K, Jerome RE, Edenberg HJ. A retroviral repetitive element confers tissue-specificity to the human alcohol dehydrogenase 1C (ADH1C) gene. DNA Cell Biol. 2002;21:793–801.

263. Beyer U, Moll-Rocek J, Moll UM, Dobbelstein M. Endogenous retrovirus drives hitherto unknown proapoptotic p63 isoforms in the male germ line of humans and great apes. Proc Natl Acad Sci U S A. 2011;108:3624–9.

264. Pi W, Zhu X, Wu M, Wang Y, Fulzele S, Eroglu A, et al. Long-range function of an intergenic retrotransposon. Proc Natl Acad Sci U S A. 2010;107:12992–7.

265. Krönung SK, Beyer U, Chiaramonte ML, Dolfini D, Mantovani R, Dobbelstein M. LTR12 promoter activation in a broad range of human tumor cells by HDAC inhibition. Oncotarget. 2016;7:33484–97.

266. Xu L, Elkahloun AG, Candotti F, Grajkowski A, Beaucage SL, Petricoin EF, et al. A novel function of RNAs arising from the long terminal repeat of human endogenous retrovirus 9 in cell cycle arrest. J Virol. 2013;87:25–36.

267. Yu X, Zhu X, Pi W, Ling J, Ko L, Takeda Y, et al. The long terminal repeat (LTR) of ERV-9 human endogenous retrovirus binds to NF-Y in the assembly of an active LTR enhancer complex NF-Y/MZF1/GATA-2. J Biol Chem. 2005;280:35184–94.

268. Lavie L, Kitova M, Maldener E, Meese E, Mayer J. CpG methylation directly regulates transcriptional activity of the human endogenous retrovirus family HERV-K(HML-2). J Virol. 2005;79:876–83.

269. Colombo AR, Elias HK, Ramsingh G. Senescence induction universally activates transposable element expression. Cell Cycle Georget Tex. 2018;17:1846–57.

270. Bardi G, Fenger C, Johansson B, Mitelman F, Heim S. Tumor karyotype predicts clinical outcome in colorectal cancer patients. J Clin Oncol Off J Am Soc Clin Oncol. 2004;22:2623–34.

271. Küppers R. Molecular biology of Hodgkin lymphoma. Hematol Am Soc Hematol Educ Program. 2009;491–6.

272. Ansell SM. Hodgkin Lymphoma: Diagnosis and Treatment. Mayo Clin Proc. 2015;90:1574–83.

273. Mathas S, Hartmann S, Küppers R. Hodgkin lymphoma: Pathology and biology. Semin Hematol. 2016;53:139–47.

274. Küppers R. The biology of Hodgkin’s lymphoma. Nat Rev Cancer. 2009;9:15–27.

275. Küppers R, Engert A, Hansmann M-L. Hodgkin lymphoma. J Clin Invest. 2012;122:3439–47.

276. Steidl C, Connors JM, Gascoyne RD. Molecular pathogenesis of Hodgkin’s lymphoma: increasing evidence of the importance of the microenvironment. J Clin Oncol Off J Am Soc Clin Oncol. 2011;29:1812–26.

277. Connors JM, Ansell SM, Fanale M, Park SI, Younes A. Five-year follow-up of brentuximab vedotin combined with ABVD or AVD for advanced-stage classical Hodgkin lymphoma. Blood. 2017;130:1375–7.

278. Mottok A, Steidl C. Biology of classical Hodgkin lymphoma: implications for prognosis and novel therapies. Blood. 2018;131:1654–65.

279. Chan FC, Mottok A, Gerrie AS, Power M, Nijland M, Diepstra A, et al. Prognostic Model to Predict Post-Autologous Stem-Cell Transplantation Outcomes in Classical Hodgkin Lymphoma. J Clin Oncol Off J Am Soc Clin Oncol. 2017;35:3722–33.

280. Boyne DJ, Mickle AT, Brenner DR, Friedenreich CM, Cheung WY, Tang KL, et al. Long-term risk of cardiovascular mortality in lymphoma survivors: A systematic review and meta-analysis. Cancer Med. 2018;7:4801–13.

281. Steinhardt JJ, Gartenhaus RB. Promising Personalized Therapeutic Options for Diffuse Large B-cell Lymphoma Subtypes with Oncogene Addictions. Clin Cancer Res. 2012;18:4538–48.

282. Boltežar L, Prevodnik VK, Perme MP, Gašljević G, Novaković BJ. Comparison of the algorithms classifying the ABC and GCB subtypes in diffuse large B-cell lymphoma. Oncol Lett. 2018;15:6903–12.

283. Reddy A, Zhang J, Davis NS, Moffitt AB, Love CL, Waldrop A, et al. Genetic and Functional Drivers of Diffuse Large B Cell Lymphoma. Cell. 2017;171:481–494.e15.

284. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, et al. The Human Genome Browser at UCSC. Genome Res. 2002;12:996–1006.

285. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012;9:357–9.

286. Schneider CA, Rasband WS, Eliceiri KW. NIH Image to ImageJ: 25 years of image analysis. Nat Methods. 2012;9:671–5.

287. Livak KJ, Schmittgen TD. Analysis of relative gene expression data using real-time quantitative PCR and the 2(-Delta Delta C(T)) Method. Methods San Diego Calif. 2001;25:402–8.

288. Rebollo R, Karimi MM, Bilenky M, Gagnier L, Miceli-Royer K, Zhang Y, et al. Retrotransposon-induced heterochromatin spreading in the mouse revealed by insertional polymorphisms. PLoS Genet. 2011;7:e1002301.

289. Perkins JR, Dawes JM, McMahon SB, Bennett DL, Orengo C, Kohl M. ReadqPCR and NormqPCR: R packages for the reading, quality checking and normalisation of RT-qPCR quantification cycle (Cq) data. BMC Genomics. 2012;13:296.

290. Brumbaugh CD, Kim HJ, Giovacchini M, Pourmand N. NanoStriDE: normalization and differential expression analysis of NanoString nCounter data. BMC Bioinformatics. 2011;12:479.

291. Lazzari E, Jefferies CA. IRF5-mediated signaling and implications for SLE. Clin Immunol. 2014;153:343–52.

292. Lamprecht B, Walter K, Kreher S, Kumar R, Hummel M, Lenze D, et al. Derepression of an endogenous long terminal repeat activates the CSF1R proto-oncogene in human lymphoma. Nat Med. 2010;16:571–579, 1p following 579.

293. Lamprecht B, Bonifer C, Mathas S. Repeat-element driven activation of proto-oncogenes in human malignancies. Cell Cycle Georget Tex. 2010;9:4276–81.

294. Edginton-White B, Cauchy P, Assi SA, Hartmann S, Riggs AG, Mathas S, et al. Global long terminal repeat activation participates in establishing the unique gene expression programme of classical Hodgkin lymphoma. Leukemia. 2018;

295. Huang Y, Zhao W. Somatic Mutations of Epigenetic Regulator Genes in Diffuse Large B-Cell Lymphoma. Blood. 2016;128:1756–1756.

296. Leung A, Trac C, Kato H, Costello KR, Chen Z, Natarajan R, et al. LTRs activated by Epstein-Barr virus-induced transformation of B cells alter the transcriptome. Genome Res. 2018;28:1791–8.

297. Kreher S, Bouhlel MA, Cauchy P, Lamprecht B, Li S, Grau M, et al. Mapping of transcription factor motifs in active chromatin identifies IRF5 as key regulator in classical Hodgkin lymphoma. Proc Natl Acad Sci U S A. 2014;111:E4513-4522.

298. Panganiban AT. Retroviral DNA integration. Cell. 1985;42:5–6.

299. Mancl ME, Hu G, Sangster-Guity N, Olshalsky SL, Hoops K, Fitzgerald-Bocarsly P, et al. Two discrete promoters regulate the alternatively spliced human interferon regulatory factor-5 isoforms. Multiple isoforms with distinct cell type-specific expression, localization, regulation, and function. J Biol Chem. 2005;280:21078–90.

300. Rice P, Longden I, Bleasby A. EMBOSS: the European Molecular Biology Open Software Suite. Trends Genet TIG. 2000;16:276–7.

301. Busam KJ, Vilain RE, Lum T, Busam JA, Hollmann TJ, Saw RPM, et al. Primary and Metastatic Cutaneous Melanomas Express ALK Through Alternative Transcriptional Initiation. Am J Surg Pathol. 2016;40:786–95.

302. Xue X, Gao W, Sun B, Xu Y, Han B, Wang F, et al. Vasohibin 2 is transcriptionally activated and promotes angiogenesis in hepatocellular carcinoma. Oncogene. 2013;32:1724–34.

303. Tu M, Li Z, Liu X, Lv N, Xi C, Lu Z, et al. Vasohibin 2 promotes epithelial-mesenchymal transition in human breast cancer via activation of transforming growth factor β 1 and hypoxia dependent repression of GATA-binding factor 3. Cancer Lett. 2017;388:187–97.

304. Norita R, Suzuki Y, Furutani Y, Takahashi K, Yoshimatsu Y, Podyma-Inoue KA, et al. Vasohibin-2 is required for epithelial-mesenchymal transition of ovarian cancer cells by modulating transforming growth factor-β signaling. Cancer Sci. 2017;108:419–26.

305. Ge Q, Zhou J, Tu M, Xue X, Li Z, Lu Z, et al. Nuclear vasohibin-2 promotes cell proliferation by inducing G0/G1 to S phase progression. Oncol Rep. 2015;34:1327–36.

306. Zhao S, Geybels MS, Leonardson A, Rubicz R, Kolb S, Yan Q, et al. Epigenome-Wide Tumor DNA Methylation Profiling Identifies Novel Prognostic Biomarkers of Metastatic-Lethal Progression in Men Diagnosed with Clinically Localized Prostate Cancer. Clin Cancer Res Off J Am Assoc Cancer Res. 2017;23:311–9.

307. Oxford G, Owens CR, Titus BJ, Foreman TL, Herlevsen MC, Smith SC, et al. RalA and RalB: antagonistic relatives in cancer cell migration. Cancer Res. 2005;65:7111–20.

308. Lim K-H, O'Hayer K, Adam SJ, Kendall SD, Campbell PM, Der CJ, et al. Divergent roles for RalA and RalB in malignant growth of human pancreatic carcinoma cells. Curr Biol CB. 2006;16:2385–94.

309. Zago G, Veith I, Singh M, Fuhrmann L, De Beco S, Remorino A, et al. RalB directly triggers invasion downstream Ras by mobilizing the Wave complex. eLife. 2018;7.

310. Pomeroy EJ, Lee LA, Lee RDW, Schirm DK, Temiz NA, Ma J, et al. Ras oncogene-independent activation of RALB signaling is a targetable mechanism of escape from NRAS(V12) oncogene addiction in acute myeloid leukemia. Oncogene. 2017;36:3263–73.

311. Yan C, Theodorescu D. RAL GTPases: Biology and Potential as Therapeutic Targets in Cancer. Pharmacol Rev. 2018;70:1–11.

312. Irsch J, Nitsch S, Hansmann M-L, Rajewsky K, Tesch H, Diehl V, et al. Isolation of viable Hodgkin and Reed-Sternberg cells from Hodgkin disease tissues. Proc Natl Acad Sci. 1998;95:10117–22.

313. Gerstein MB, Bruce C, Rozowsky JS, Zheng D, Du J, Korbel JO, et al. What is a gene, post-ENCODE? History and updated definition. Genome Res. 2007;17:669–81.

314. Waddington CH. Genetic Assimilation of the Bithorax Phenotype. Evolution. 1956;10:1–13.

315. Jablonka E, Lamb MJ. The inheritance of acquired epigenetic variations. J Theor Biol. 1989;139:69–83.

316. Berger SL, Kouzarides T, Shiekhattar R, Shilatifard A. An operational definition of epigenetics. Genes Dev. 2009;23:781–3.

317. Molaro A, Malik HS. Hide and seek: how chromatin-based pathways silence retroelements in the mammalian germline. Curr Opin Genet Dev. 2016;37:51–8.

318. Tsumura A, Hayakawa T, Kumaki Y, Takebayashi S, Sakaue M, Matsuoka C, et al. Maintenance of self-renewal ability of mouse embryonic stem cells in the absence of DNA methyltransferases Dnmt1, Dnmt3a and Dnmt3b. Genes Cells Devoted Mol Cell Mech. 2006;11:805–14.

319. Liu S, Brind'Amour J, Karimi MM, Shirane K, Bogutz A, Lefebvre L, et al. Setdb1 is required for germline development and silencing of H3K9me3-marked endogenous retroviruses in primordial germ cells. Genes Dev. 2014;28:2041–55.

320. Castro-Diaz N, Ecco G, Coluccio A, Kapopoulou A, Yazdanpanah B, Friedli M, et al. Evolutionally dynamic L1 regulation in embryonic stem cells. Genes Dev [Internet]. 2014 [cited 2018 Oct 30]; Available from: http://genesdev.cshlp.org/content/early/2014/06/17/gad.241661.114

321. Walter M, Teissandier A, Pérez-Palacios R, Bourc'his D. An epigenetic switch ensures transposon repression upon dynamic loss of DNA methylation in embryonic stem cells. Ferguson-Smith AC, editor. eLife. 2016;5:e11418.

322. Orr WC. Tightening the connection between transposable element mobilization and aging. Proc Natl Acad Sci U S A. 2016;113:11069–70.

323. Bollati V, Schwartz J, Wright R, Litonjua A, Tarantini L, Suh H, et al. Decline in genomic DNA methylation through aging in a cohort of elderly subjects. Mech Ageing Dev. 2009;130:234–9.

324. Horvath S, Zhang Y, Langfelder P, Kahn RS, Boks MP, van Eijk K, et al. Aging effects on DNA methylation modules in human brain and blood tissue. Genome Biol. 2012;13:R97.

325. Laurent GS, Hammell N, McCaffrey TA. A LINE-1 Component to Human Aging: Do LINE elements exact a longevity cost for evolutionary advantage? Mech Ageing Dev. 2010;131:299–305.

326. Mustafina OE. The possible roles of human Alu elements in aging. Front Genet [Internet]. 2013 [cited 2018 Oct 30];4. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3664780/

327. Sharma S, Kelly TK, Jones PA. Epigenetics in cancer. Carcinogenesis. 2010;31:27–36.

328. Morin RD, Johnson NA, Severson TM, Mungall AJ, An J, Goya R, et al. Somatic mutations altering EZH2 (Tyr641) in follicular and diffuse large B-cell lymphomas of germinal-center origin. Nat Genet. 2010;42:181–5.

329. Ley TJ, Ding L, Walter MJ, McLellan MD, Lamprecht T, Larson DE, et al. DNMT3A Mutations in Acute Myeloid Leukemia. N Engl J Med. 2010;363:2424–33.

330. Shain AH, Pollack JR. The spectrum of SWI/SNF mutations, ubiquitous in human cancers. PloS One. 2013;8:e55119.

331. Mack SC, Witt H, Piro RM, Gu L, Zuyderduyn S, Stütz AM, et al. Epigenomic alterations define lethal CIMP-positive ependymomas of infancy. Nature. 2014;506:445–50.

332. Böhne A, Brunet F, Galiana-Arnoux D, Schultheis C, Volff J-N. Transposable elements as drivers of genomic and biological diversity in vertebrates. Chromosome Res Int J Mol Supramol Evol Asp Chromosome Biol. 2008;16:203–15.

333. Rebollo R, Horard B, Hubert B, Vieira C. Jumping genes and epigenetics: Towards new species. Gene. 2010;454:1–7.

334. Friedli M, Trono D. The developmental control of transposable elements and the evolution of higher species. Annu Rev Cell Dev Biol. 2015;31:429–51.

335. Warren IA, Naville M, Chalopin D, Levin P, Berger CS, Galiana D, et al. Evolutionary impact of transposable elements on genomic diversity and lineage-specific innovation in vertebrates. Chromosome Res Int J Mol Supramol Evol Asp Chromosome Biol. 2015;23:505–31.

336. Hansen KD, Timp W, Bravo HC, Sabunciyan S, Langmead B, McDonald OG, et al. Increased methylation variation in epigenetic domains across cancer types. Nat Genet. 2011;43:768–75.

337. Landau DA, Clement K, Ziller MJ, Boyle P, Fan J, Gu H, et al. Locally disordered methylation forms the basis of intratumor methylome variation in chronic lymphocytic leukemia. Cancer Cell. 2014;26:813–25.

338. Li S, Garrett-Bakelman FE, Chung SS, Sanders MA, Hricik T, Rapaport F, et al. Distinct evolution and dynamics of epigenetic and genetic heterogeneity in acute myeloid leukemia. Nat Med. 2016;22:792–9.

339. Mazor T, Pankov A, Song JS, Costello JF. Intratumoral Heterogeneity of the Epigenome. Cancer Cell. 2016;29:440–51.

340. Timp W, Feinberg AP. Cancer as a dysregulated epigenome allowing cellular growth advantage at the expense of the host. Nat Rev Cancer. 2013;13:497–510.

341. Shen H, Laird PW. Interplay between the Cancer Genome and Epigenome. Cell. 2013;153:38–55.

342. Berdasco M, Esteller M. Aberrant epigenetic landscape in cancer: how cellular identity goes awry. Dev Cell. 2010;19:698–711.

343. Nephew KP, Huang TH-M. Epigenetic gene silencing in cancer initiation and progression. Cancer Lett. 2003;190:125–33.

344. Kazanets A, Shorstova T, Hilmi K, Marques M, Witcher M. Epigenetic silencing of tumor suppressor genes: Paradigms, puzzles, and potential. Biochim Biophys Acta. 2016;1865:275–88.

345. Brock A, Chang H, Huang S. Non-genetic heterogeneity--a mutation-independent driving force for the somatic evolution of tumours. Nat Rev Genet. 2009;10:336–42.

346. Werfel J, Krause S, Bischof AG, Mannix RJ, Tobin H, Bar-Yam Y, et al. How changes in extracellular matrix mechanics and gene expression variability might combine to drive cancer progression. PloS One. 2013;8:e76122.

347. Pisco AO, Brock A, Zhou J, Moor A, Mojtahedi M, Jackson D, et al. Non-Darwinian dynamics in therapy-induced cancer drug resistance. Nat Commun. 2013;4:2467.

348. Marusyk A, Almendro V, Polyak K. Intra-tumour heterogeneity: a looking glass for cancer? Nat Rev Cancer. 2012;12:323–34.

349. Bashtrykov P, Jankevicius G, Smarandache A, Jurkowska RZ, Ragozin S, Jeltsch A. Specificity of Dnmt1 for methylation of hemimethylated CpG sites resides in its catalytic domain. Chem Biol. 2012;19:572–8.

350. Jang HS, Shah NM, Du AY, Dailey ZZ, Pehrsson EC, Godoy PM, et al. Transposable elements drive widespread expression of oncogenes in human cancers. Nat Genet. 2019;51:611.

351. Coffin JM, Hughes SH, Varmus HE. Genetic Organization [Internet]. Cold Spring Harbor Laboratory Press; 1997 [cited 2019 Jan 22]. Available from: https://www.ncbi.nlm.nih.gov/books/NBK19370/

352. Clark DN, Read RD, Mayhew V, Petersen SC, Argueta LB, Stutz LA, et al. Four Promoters of IRF5 Respond Distinctly to Stimuli and are Affected by Autoimmune-Risk Polymorphisms. Front Immunol. 2013;4:360.

353. Chiaromonte F, Yap VB, Miller W. Scoring pairwise genomic sequence alignments. Pac Symp Biocomput Pac Symp Biocomput. 2002;115–26.

Appendices

A: Supplementary materials: Chapter 2

Name	Classification	Type	LibraryID	Reference	Read Length
GM12878	B-lymphoblastoid	Cell Line	wgEncodeEH000122	ENCODE [9]	2 x 75 nt
K562	Chronic Myelogenous Leukemia	Cell Line	wgEncodeEH000126	ENCODE [9]	2 x 75 nt
H1 esc	Human embryonic stem cell	Cell Line	wgEncodeEH000128	ENCODE [9]	2 x 75 nt
L428	classical HL	Cell Line	HS0999	Steidl et al, 2011 [230]	2 x 50 nt
L540	classical HL	Cell Line	A05247	Liu et al., 2014 [231]	2 x 75 nt
L591	classical HL	Cell Line	A05250	Babaian et al., 2016 [213]	2 x 75 nt
L1236	classical HL	Cell Line	A05254	Liu et al., 2014 [231]	2 x 75 nt
HDLM-2	classical HL	Cell Line	A05248	Babaian et al., 2016 [213]	2 x 75 nt
KM-H2	classical HL	Cell Line	HS0988	Steidl et al, 2011 [230]	2 x 50 nt
SUP-HD1	classical HL	Cell Line	A05251	Liu et al., 2014 [231]	2 x 75 nt
U-H01	classical HL	Cell Line	A05249	Babaian et al., 2016 [213]	2 x 75 nt
DEV	Nodular Lymphocyte Predominant HL	Cell Line	HS2171	Twa et al, 2014 [232]	2 x 50 nt
Karpas1106p	Primary Mediastinal Large B-Cell Lymphoma	Cell Line	HS1484	Babaian et al., 2016 [213]	2 x 50 nt
MEDB1	Primary Mediastinal Large B-Cell Lymphoma	Cell Line	A05253	Babaian et al., 2016 [213]	2 x 75 nt
U2940	Primary Mediastinal Large B-Cell Lymphoma	Cell Line	A05252	Babaian et al., 2016 [213]	2 x 75 nt
B-Cell 1	B-cell	Primary	HS0669	Morin et al., 2011 [169]	2 x 36 nt
B-Cell 2	B-cell	Primary	HS0670	Morin et al., 2011 [169]	2 x 50 nt
B-Cell 3	B-cell	Primary	HS1044	Morin et al., 2011 [169]	2 x 50 nt
B-Cell 4	B-cell	Primary	HS1045	Morin et al., 2011 [169]	2 x 50 nt
B-Cell 5	B-cell	Primary	HS2253	Morin et al., 2011 [169]	2 x 75 nt
B-Cell 6	B-cell	Primary	HS2254	Morin et al., 2011 [169]	2 x 75 nt
B-Cell 7	B-cell	Primary	HS2639	Morin et al., 2011 [169]	2 x 75 nt
B-Cell 8	B-cell	Primary	HS2640	Morin et al., 2011 [169]	2 x 75 nt
B-Cell 9	B-cell	Primary	HS2641	Morin et al., 2011 [169]	2 x 75 nt

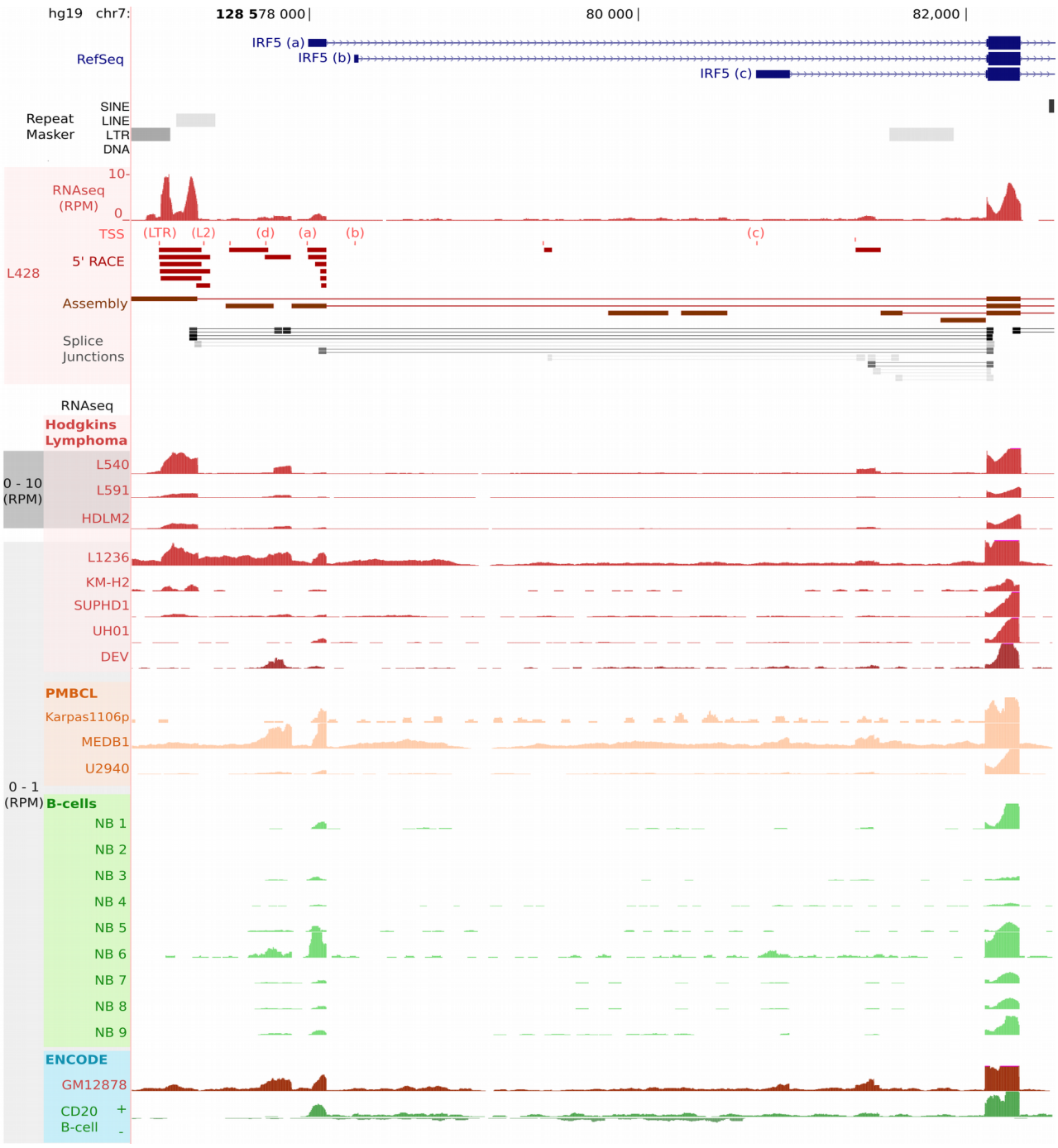
Supplementary Table 2.1: RNA-seq data-sets

{Update References}

Pair #	Primer Name	Sequence	Tm	size
1	NLRP1 ex1 LTR-s	CTGAACTGCGCTGTTCTTGC	66	~1000
1	NLRP1 ex4as	GCCTGCCTTTCTCTGATTTC	63	
2	DHRS2-LTR-F	GCAGTGAGACTATTGCCAAGTG	64	
2	DHRS2-ex3-R	GAAAGCCTAGCACAGGGATG	64	168
3	IL1R2 LTR-F	GACGCTCATACAAATCAACAG	60	114
3	IL1R2 Native-R	TGACAACTTCCAGAGGACAC	61	
4	TPRG1-MIRc-s	CATCCAGCTCACTGCACTTT	63	480
4	TPRG1-ex6-as	AATAGCGTGGGTCAAACTGG	64	
5	ANKRD44-LTR-s	GGCTTCCCCTTCACTTTCTG	65	300, 400
5	ANKRD44-ex?-as	AGGCAGGGTGTTTTCAACTG	64	
6	NAALADL2-LTR-as	CCATCTGCCTTGATTGTGAG	64	115
6	NAALADL2-ex?-s	TCCCTGAGGAATTTCACCAA	65	
7	RNF19A-L2-s	ACAGCCTCTTTGGTTTCTGTT	62	754
7	RNF19A-ex2-as	ACAAGCCACCGTCTAAGCAT	64	
8	PIP5K1B-s	CCCAGTTACTTGGGAGCGTA	64	173
8	PIP5K1B-MLT-as	AGAGGGAAAACCCTGCTGAT	64	
9	FHAD1-LTRs	ATTGCTGAGGAGCCAGAGAG	64	215
9	FHAD1-ex3-as	TTCAATGAGTGCATGGTGGT	64	
10	NCF2-LTR-s	TTCATTTGGGACCAGTAGCC	64	327
10	NCF2-ex2-as	GCATCCCTCGTTGGAAGTAA	64	
11	C1orf186-LTR-s	ATTTGGTGTCTGAGGGGTTTT	64	330
11	C1orf186-ex?-as	GCTTCAGGGTGGTGATGTTC	65	
12	VASH2-MLT-s	GCATGGGACTTCTTGACCTC	64	580
12	VASH2-ex2-as	TCTTGTTGACGTGGAACAGC	64	
13	HBE1-ex1-as	GAGGGTCAGCAGTGATGGAT	64	156
13	HBE1-LTR-s	GCCATTCCAGTAGGATGTGA	63	

Supplementary Table 2.2: RT-PCR Primers

B: Supplementary materials: Chapter 4



Supplementary Figure 4.1: IRF5 promoter in cHL and B-cell controls

index_contigID	Exon #	Gene Symbol	TE-Gene Overlap	TE Name	Repeat-Exon Coordinate	Interaction Type	# of Norm.	# of Canc.	Library ID
L540.1391.2	12	FRRS1; PALMD	c	MIRb:SINE:MIR	chr1:100214129-100214595	Up	0	9	L540;HDLM2;UH01;L591;SUPHD1;U2940_;MEDB1_;L1236;L428
HDLM2.15904.2	20	LOC728392; NLRP1	s	THE1C:LTR:ERVL-MaLR	chr17:5487613-5522932	Up	0	9	HDLM2;UH01;SUPHD1;U2940_;L1236;KMH2;L428;Karpas1106p_;DEV
L540.8069.2	1	DHRS2	s	LTR12D:LTR:ERV1	chr14:24104835-24105981	UpEdge	0	8	L540;HDLM2;UH01;SUPHD1;U2940_;L1236;KMH2;L428
L540.16605.1	1	IL1R2	s	MLT1H1:LTR:ERVL-MaLR	chr2:102614908-102615578	UpEdge	0	8	L540;HDLM2;L591;SUPHD1;MEDB1_;L1236;KMH2;L428
UH01.28016.1	1	APOL2; APOL3; APOL4	s	LTR57:LTR:ERVL	chr22:36621955-36663652	Up	0	8	UH01;SUPHD1;U2940_;MEDB1_;L1236;KMH2;Karpas1106p_;DEV
L540.21824.2	1	TPRG1	s	MIRc:SINE:MIR	chr3:189025974-189027966	Up	0	8	L540;HDLM2;UH01;L591;U2940_;MEDB1_;L1236;KMH2r
HDLM2.9660.1	1	.	i	L2:LINE:L2	chr12:92950830-92951460	EInside	0	7	HDLM2;UH01;L591;SUPHD1;MEDB1_;L1236;L428
HDLM2.11405.2	3	.	i	THE1A:LTR:ERVL-MaLR	chr14:21132287-21135138	UpEdge	0	7	HDLM2;L591;SUPHD1;U2940_;MEDB1_;L1236;DEV
HDLM2.23472.2	28	ANKRD44; .	s	THE1D:LTR:ERVL-MaLR	chr2:198220738-198221091	EInside	0	7	HDLM2;UH01;SUPHD1;U2940_;L428;Karpas1106p_;DEV
U2940_.34764.1	1	NAALADL2	as	THE1B:LTR:ERVL-MaLR	chr3:174943362-174992161	Up	0	7	U2940_;MEDB1_;L1236;KMH2;L428;Karpas1106p_;DEV
L591.27147.1	2	.	i	THE1B:LTR:ERVL-MaLR	chr9:1929260-1930573	Up	0	7	L591;U2940_;MEDB1_;KMH2;L428;Karpas1106p_;DEV
L540.710.2	2	THRAP3	s	MIR:SINE:MIR	chr1:36690961-36725085	Up	0	6	L540;UH01;SUPHD1;U2940_;MEDB1_;L1236
L540.3843.1	1	BBIP1; PDCD4	c	MER103C:DNA:hAT-Charlie	chr10:112628632-112631776	UpEdge	0	6	L540;HDLM2;UH01;L591;SUPHD1;MEDB1_
L540.6864.2	1	.	i	MLT1F:LTR:ERVL-MaLR	chr12:104319102-104319588	EInside	0	6	L540;HDLM2;L591;SUPHD1;KMH2;Karpas1106p_
L540.7945.2	2	.	i	L1ME2:LINE:L1	chr13:110773934-110774690	Up	0	6	L540;UH01;L591;SUPHD1;L1236;L428
UH01.12570.1	3	.	i	MLT1E2:LTR:ERVL-MaLR	chr14:94456091-94456562	EInside	0	6	UH01;L591;MEDB1_;L428;Karpas1106p_;DEV
L540.23128.3	3	.	i	MER50:LTR:ERV1	chr4:161480550-161747548	Up	0	6	L540;HDLM2;UH01;MEDB1_;L1236;L428
UH01.39272.3	22	TNPO3; IRF5	c	L1MB7:LINE:L1	chr7:128696049-128699202	UpEdge	0	6	UH01;SUPHD1;U2940_;MEDB1_;L1236;L428
L540.27107.2	2	.	i	MIR3:SINE:MIR	chr7:45034628-45039013	Up	0	6	L540;HDLM2;UH01;L591;SUPHD1;MEDB1_
L540.29119.1	9	RNF19A	s	L2a:LINE:L2	chr8:101299729-101326125	Up	0	6	L540;HDLM2;UH01;SUPHD1;MEDB1_;L1236
HDLM2.38711.1	1	.	i	MER61A:LTR:ERV1	chr9:31848548-31848905	UpEdge	0	6	HDLM2;UH01;L591;SUPHD1;KMH2;L428
L540.29918.1	1	PIP5K1B	as	MLT1C:LTR:ERVL-MaLR	chr9:71571949-71590955	Up	0	6	L540;HDLM2;UH01;L591;SUPHD1;L428
L540.254.1	1	FHAD1	s	MLT1K:LTR:ERVL-MaLR	chr1:15562768-15563305	EInside	0	5	L540;HDLM2;UH01;L1236;L428
L591.2361.1	16	NCF2; SMG7	c	LTR27B:LTR:ERV1	chr1:183559758-183560407	UpEdge	0	5	L591;MEDB1_;L428;Karpas1106p_;DEV
HDLM2.3360.1	4	C1orf186; ANKRD65	s	Harlequin-int:LTR:ERV1	chr1:206284908-206288369	UpEdge	0	5	HDLM2;UH01;L591;SUPHD1;KMH2r
HDLM2.3492.3	1	VASH2	s	MLT2B2:LTR:ERVL	chr1:213104236-213104759	EInside	0	5	HDLM2;L591;MEDB1_;L1236;L428
SUPHD1.8499.1	15	MDM1; .	s	THE1A:LTR:ERVL-MaLR	chr12:68726243-68836118	Up	0	5	SUPHD1;L1236;KMH2;L428;DEV
HDLM2.10775.1	1	.	i	LTR2:LTR:ERV1	chr13:41444892-41455418	UpEdge	0	5	HDLM2;SUPHD1;U2940_;MEDB1_;L1236
L540.8490.2	1	FUT8	s	L2a:LINE:L2	chr14:65802821-65803573	EInside	0	5	L540;HDLM2;SUPHD1;L1236;L428
UH01.19418.1	1	.	i	MLT1M:LTR:ERVL-MaLR	chr18:48918183-48918627	UpEdge	0	5	UH01;L591;MEDB1_;KMH2;L428
UH01.29429.2	3	.; PRICKLE2	s	L3:LINE:CR1	chr3:64444966-64448086	Up	0	5	UH01;L591;L1236;Karpas1106p_;DEV
L540.22918.1	1	AP1AR; TIFA; ALPK1	c	L1ME4a:LINE:L1	chr4:113150608-113153330	UpEdge	0	5	L540;HDLM2;UH01;L591;L1236
HDLM2.31531.3	1	ARL14EPL	s	MER39B:LTR:ERV1	chr5:115383199-115383702	EInside	0	5	HDLM2;SUPHD1;MEDB1_;L1236;L428
UH01.39649.1	1	.	i	MIRc:SINE:MIR	chr7:150101438-150105576	UpEdge	0	5	UH01;L591;U2940_;L1236;KMH2r
HDLM2.35031.2	2	.; GARS	as	LOR1-int:LTR:ERV1	chr7:30618331-30618869	EInside	0	5	HDLM2;UH01;L591;SUPHD1;MEDB1_
UH01.38217.1	1	.	i	MLT2D:LTR:ERVL	chr7:45763397-45764023	UpEdge	0	5	UH01;L591;SUPHD1;U2940_;MEDB1_
UH01.38758.2	2	.	i	THE1C:LTR:ERVL-MaLR	chr7:93651925-93690937	Up	0	5	UH01;L591;U2940_;L428;Karpas1106p_
UH01.41332.1	1	.	i	MLT1I:LTR:ERVL-MaLR	chr8:140613078-140614513	Up	0	5	UH01;MEDB1_;L428;Karpas1106p_;DEV
HDLM2.39761.1	1	OR1J1; .	as	L1PB1:LINE:L1	chr9:125223662-125227028	UpEdge	0	5	HDLM2;UH01;SUPHD1;L1236;L428
L540.196.1	1	.	u	AluSx:SINE:Alu	chr1:10452252-155111263	Up	0	4	L540;UH01;SUPHD1;MEDB1_
SUPHD1.1729.3	4	ADORA3	s	AluY:SINE:Alu	chr1:112031299-112040030	Up	0	4	SUPHD1;U2940_;L428;Karpas1106p_
L591.2398.1	1	.	i	MER21B:LTR:ERVL	chr1:200380598-200381432	EInside	0	4	L591;L1236;KMH2;Karpas1106p_
L540.1153.1	1	.	i	THE1B:LTR:ERVL-MaLR	chr1:69843100-69848259	Up	0	4	L540;HDLM2;UH01;KMH2r
HDLM2.4485.1	2	.	i	HERVE-int:LTR:ERV1	chr10:15022523-15036886	Up	0	4	HDLM2;UH01;SUPHD1;L1236
L540.5533.1	1	KIRREL3	as	MSTA:LTR:ERVL-MaLR	chr11:126517837-126518201	EInside	0	4	L540;HDLM2;L591;KMH2r
HDLM2.6441.1	2	ANO3	s	L1P3b:LINE:L1	chr11:26628092-26646009	Up	0	4	HDLM2;SUPHD1;L1236;L428
L540.4324.1	4	BBOX1; .	as	LTR33:LTR:ERVL	chr11:27238864-27255745	Up	0	4	L540;HDLM2;L591;SUPHD1

continued on next page...

index_contigID	Exon #	Gene Symbol	TE-Gene Overlap	TE Name	Repeat-Exon Coordinate	Interaction Type	# of Norm.	# of Canc.	Library ID
L540.5862.1	1	.; LOH12CR1	as	MIRb:SINE:MIR	chr12:12508346-12509809	UpEdge	0	4	L540;UH01;SUPHD1;L428
L540.7317.3	9	ANKLE2	s	AluSg4:SINE:Alu	chr12:133333220-133333546	UpEdge	0	4	L540;SUPHD1;U2940_;MEDB1_
HDLM2.8616.1	1	.	i	LTR12F:LTR:ERV1	chr12:14369477-14369727	UpEdge	0	4	HDLM2;MEDB1_;L1236;KMH2r
UH01.8267.1	2	.	i	THE1B:LTR:ERVL-MaLR	chr12:25108302-25113029	Up	0	4	UH01;SUPHD1;KMH2;L428
UH01.8266.1	1	.; BCAT1	s	MER4A:LTR:ERV1	chr12:25108465-25116894	Up	0	4	UH01;KMH2;L428;DEV
L540.7327.1	2	.	i	LTR79:LTR:ERVL	chr13:19295339-19301740	Up	0	4	L540;HDLM2;SUPHD1;L1236
L540.7673.1	1	FNDC3A	s	L2b:LINE:L2	chr13:49546825-49550313	UpEdge	0	4	L540;HDLM2;UH01;L1236
L591.8694.1	2	ANKRD9	s	MIR3:SINE:MIR	chr14:102974812-102975225	Up	0	4	L591;SUPHD1;L1236;L428
UH01.12279.1	1	.	i	MER52D:LTR:ERV1	chr14:71697497-71699966	UpEdge	0	4	UH01;MEDB1_;L1236;DEV
L540.8594.1	2	.	i	MIR:SINE:MIR	chr14:74296599-74297041	UpEdge	0	4	L540;HDLM2;L591;SUPHD1
UH01.13119.1	14	ARHGAP11B; .	s	Tigger1:DNA:TcMar-Tigger	chr15:31059631-31065199	UpEdge	0	4	UH01;L591;SUPHD1;MEDB1_
L591.9156.1	1	UNC13C	s	MER73:LTR:ERVL	chr15:54875840-54876417	EInside	0	4	L591;KMH2;Karpas1106p_;DEV
HDLM2.15616.1	1	.	i	MIRb:SINE:MIR	chr16:85337091-85337307	EInside	0	4	HDLM2;UH01;SUPHD1;L1236
L591.13933.1	4	ZNF566	s	SVA_D:Other:Other	chr19:36980388-36983060	UpEdge	0	4	L591;U2940_;L428;DEV
L540.14867.2	5	ZNF45	s	MIR:SINE:MIR	chr19:44428390-44428582	UpEdge	0	4	L540;HDLM2;L591;U2940_
SUPHD1.18693.4	3	BBC3	s	MIR3:SINE:MIR	chr19:47730185-47730457	UpEdge	0	4	SUPHD1;U2940_;MEDB1_;L428
SUPHD1.18883.2	4	C19orf48	s	AluSg:SINE:Alu	chr19:51305711-51306042	UpEdge	0	4	SUPHD1;U2940_;MEDB1_;L1236
L540.16775.1	1	.	i	MLT1A0:LTR:ERVL-MaLR	chr2:127984704-127985031	EInside	0	4	L540;L591;SUPHD1;L1236
L540.17413.1	1	.	i	LTR12C:LTR:ERV1	chr2:191618534-191626254	Up	0	4	L540;HDLM2;SUPHD1;L1236
L540.17431.1	1	.	i	AluJb:SINE:Alu	chr2:194745497-194758883	Up	0	4	L540;L591;SUPHD1;L428
L540.15543.1	1	.	i	MER44C:DNA:TcMar-Tigger	chr2:7862414-7863826	Up	0	4	L540;HDLM2;SUPHD1;L428
L540.15564.1	3	.	i	THE1C:LTR:ERVL-MaLR	chr2:8071935-8073376	Up	0	4	L540;HDLM2;L1236;L428
HDLM2.24950.1	1	CYYR1; .	as	ERV3-16A3_LTR:LTR:ERVL	chr21:27794516-27794878	UpEdge	0	4	HDLM2;UH01;SUPHD1;L1236
UH01.27600.1	1	UPB1	s	HAL1:LINE:L1	chr22:24899078-24899562	UpEdge	0	4	UH01;L591;MEDB1_;KMH2r
HDLM2.25848.1	2	MB; .	s	MLT2B1:LTR:ERVL	chr22:36006931-36031393	Up	0	4	HDLM2;L591;SUPHD1;L1236
L540.19952.2	5	RAC2; .	s	L1ME1:LINE:L1	chr22:37628841-37644487	Up	0	4	L540;HDLM2;L591;L428
UH01.28742.1	13	DYNC1LI1	s	AluJb:SINE:Alu	chr3:32612117-32614003	UpEdge	0	4	UH01;SUPHD1;KMH2;L428
L540.24313.3	3	.	i	HAL1:LINE:L1	chr5:133767963-133768793	EInside	0	4	L540;HDLM2;SUPHD1;MEDB1_
UH01.34347.1	6	GNPDA1	s	MIRb:SINE:MIR	chr5:141391477-141391992	Up	0	4	UH01;U2940_;MEDB1_;L1236
L540.24008.1	2	.	i	LTR78:LTR:ERV1	chr5:91745927-91794061	Up	0	4	L540;HDLM2;SUPHD1;L1236
L540.24008.1	3	.	i	LTR78:LTR:ERV1	chr5:91793871-91794820	UpEdge	0	4	L540;HDLM2;SUPHD1;L1236
HDLM2.31215.1	1	.	i	L1PA13:LINE:L1	chr5:96685648-96692076	EInside	0	4	HDLM2;L591;U2940_;DEV
U2940_U2940_.41192	2	.	i	L2:LINE:L2	chr6:107197505-107214030	Up	0	4	U2940_;MEDB1_;KMH2;Karpas1106p_
UH01.36732.2	1	.	i	MER52C:LTR:ERV1	chr6:113201027-113202040	UpEdge	0	4	UH01;U2940_;MEDB1_;L1236
UH01.36757.1	4	.	i	MLT1G:LTR:ERVL-MaLR	chr6:114194159-114194736	UpEdge	0	4	UH01;L591;MEDB1_;L1236
L540.26227.1	1	.	i	LTR12F:LTR:ERV1	chr6:115319218-115319583	UpEdge	0	4	L540;UH01;SUPHD1;KMH2r
HDLM2.36032.2	1	.	i	MER41B:LTR:ERV1	chr7:106415072-106415716	EInside	0	4	HDLM2;SUPHD1;L1236;L428
HDLM2.36431.1	14	SLC37A3	s	L2a:LINE:L2	chr7:140082238-140090963	Up	0	4	HDLM2;UH01;SUPHD1;L428
L540.28348.2	9	PTPRN2	s	MLT1K:LTR:ERVL-MaLR	chr7:158109511-158224794	Up	0	4	L540;L591;SUPHD1;L1236
HDLM2.35212.3	7	.	i	LTR84b:LTR:ERVL	chr7:45791128-45796598	Up	0	4	HDLM2;U2940_;MEDB1_;L1236
HDLM2.37118.1	5	FUT10	s	AluJo:SINE:Alu	chr8:33330582-33331272	UpEdge	0	4	HDLM2;UH01;L591;MEDB1_
HDLM2.37389.1	1	.	i	MER57B1:LTR:ERV1	chr8:66783519-66784040	UpEdge	0	4	HDLM2;U2940_;MEDB1_;Karpas1106p_
L540.30043.2	6	.	i	L1PB1:LINE:L1	chr9:93762545-93785950	Up	0	4	L540;HDLM2;SUPHD1;MEDB1_
HDLM2.2442.1	1	NBPF8	as	L1MEc:LINE:L1	chr1:147751151-147751933	EInside	0	3	HDLM2;SUPHD1;L1236
HDLM2.2592.1	1	THEM5; C2CD4D	s	LTR61:LTR:ERV1	chr1:151822727-151828865	Up	0	3	HDLM2;SUPHD1;L1236
HDLM2.3102.1	1	.	i	THE1A:LTR:ERVL-MaLR	chr1:180527904-180528269	EInside	0	3	HDLM2;MEDB1_;KMH2r
HDLM2.3356.2	3	RGS1; RGS13; .	as	MER51A:LTR:ERV1	chr1:192596471-192709096	Up	0	3	HDLM2;MEDB1_;L1236

continued on next page...

index_contigID	Exon #	Gene Symbol	TE-Gene Overlap	TE Name	Repeat-Exon Coordinate	Interaction Type	# of Norm.	# of Canc.	Library ID
L540.2328.1	6	ZNF281; .	s	MER5B:DNA:hAT-Charlie	chr1:200452230-200452783	UpEdge	0	3	L540;HDLM2;L1236
MEDB1_MEDB1_.398	13	TAF1A; .	s	L1MEf:LINE:L1	chr1:222765501-222766948	UpEdge	0	3	MEDB1_;L1236;KMH2r
SUPHD1.3542.1	1	SPRTN; EXOC8	c	L1M5:LINE:L1	chr1:231464018-231474350	UpEdge	0	3	SUPHD1;U2940_;L1236
L591.2842.3	1	.	i	L1ME4a:LINE:L1	chr1:234813804-234818828	Up	0	3	L591;U2940_;L1236
L540.2914.1	1	.; OR2T3	as	L1MA2:LINE:L1	chr1:248630737-248744443	Up	0	3	L540;L591;MEDB1_
UH01.1524.2	8	LRRC41; UQCRH	c	LTR2B:LTR:ERV1	chr1:46797497-46798229	UpEdge	0	3	UH01;MEDB1_;L1236
U2940_U2940_.5497.1	1	SYT15	u	L1MEg:LINE:L1	chr10:46952237-88969997	Up	0	3	U2940_;MEDB1_;L1236
UH01.4668.2	3	.	i	L1ME2:LINE:L1	chr10:49882273-49883213	UpEdge	0	3	UH01;MEDB1_;L1236
L540.3277.1	1	.	i	ERVL-E-int:LTR:ERVL	chr10:52387064-52387584	EInside	0	3	L540;HDLM2;SUPHD1
L540.3525.1	1	.	i	AluJo:SINE:Alu	chr10:75471290-75478131	Up	0	3	L540;L591;SUPHD1
SUPHD1.4720.1	1	.	i	LTR2B:LTR:ERV1	chr10:85926378-85932221	UpEdge	0	3	SUPHD1;U2940_;DEV
HDLM2.7909.1	1	.	i	LTR85b:LTR:Gypsy?	chr11:123173373-123174018	EInside	0	3	HDLM2;UH01;L591
HDLM2.8120.1	1	ACAD8; GLB1L3	c	L1MD:LINE:L1	chr11:134127993-134145264	Up	0	3	HDLM2;UH01;L591
SUPHD1.6686.1	1	.	i	MSTA:LTR:ERVL-MaLR	chr11:89984184-90094790	Up	0	3	SUPHD1;Karpas1106p_;DEV
UH01.9583.2	7	.; ACTR6	as	HERVK22-int:LTR:ERVK	chr12:100553550-100556775	UpEdge	0	3	UH01;MEDB1_;L1236
UH01.10123.2	1	.; MLXIP	s	MLT1K:LTR:ERVL-MaLR	chr12:122502048-122502588	UpEdge	0	3	UH01;L591;KMH2r
KMH2r.6649.1	2	GUCY2C	as	L2b:LINE:L2	chr12:14818180-14818949	Up	0	3	KMH2;L428;DEV
UH01.8446.1	20	CPNE8	s	MLT1C:LTR:ERVL-MaLR	chr12:39299239-39303441	Up	0	3	UH01;SUPHD1;L428
L1236.13271.1	1	.	i	L1MC1:LINE:L1	chr13:84547319-84567871	Up	0	3	L1236;KMH2;DEV
HDLM2.12242.1	1	DHRS4L2	u	L1PA6:LINE:L1	chr14:24442347-24489095	Up	0	3	HDLM2;SUPHD1;MEDB1_
L591.8274.1	7	DHRS7; .	s	THE1B:LTR:ERVL-MaLR	chr14:60631895-60677440	Up	0	3	L591;SUPHD1;KMH2r
SUPHD1.10767.1	1	SLC38A6	u	MLT1B:LTR:ERVL-MaLR	chr14:61538226-61550493	Up	0	3	SUPHD1;L1236;L428
U2940_U2940_.14621	1	BTBD7; UNC79	c	MLT1L:LTR:ERVL-MaLR	chr14:93785300-93791258	UpEdge	0	3	U2940_;L1236;L428
L591.8892.2	11	FSIP1	s	L3:LINE:CR1	chr15:40068600-40069722	Up	0	3	L591;SUPHD1;MEDB1_
L591.9123.1	1	.; MAPK6	s	THE1D:LTR:ERVL-MaLR	chr15:52295415-52295825	UpEdge	0	3	L591;L1236;L428
L540.9587.2	3	.	i	MIR3:SINE:MIR	chr15:62117490-62131861	Up	0	3	L540;HDLM2;SUPHD1
UH01.14161.1	1	.	i	L1MCc:LINE:L1	chr15:85855725-85856915	EInside	0	3	UH01;U2940_;MEDB1_
L540.9988.2	1	.	i	MLT1H2:LTR:ERVL-MaLR	chr15:89584141-89584574	EInside	0	3	L540;MEDB1_;L1236
UH01.15816.1	7	SMPD3	s	MLT1I:LTR:ERVL-MaLR	chr16:68404762-68415669	Up	0	3	UH01;L1236;L428
HDLM2.15547.2	1	.	i	MER57C2:LTR:ERV1	chr16:76613229-76613713	UpEdge	0	3	HDLM2;L591;DEV
SUPHD1.14171.2	1	ZNF778	s	L1M4:LINE:L1	chr16:89281238-89284318	UpEdge	0	3	SUPHD1;U2940_;L1236
UH01.16729.2	1	TRIM16	s	AluSx3:SINE:Alu	chr17:15531203-18639263	Up	0	3	UH01;L591;L1236
HDLM2.16755.1	1	HSD17B1; .	as	MER21A:LTR:ERVL	chr17:40696911-40704578	Up	0	3	HDLM2;U2940_;KMH2r
L540.13225.2	1	.	i	LTR12C:LTR:ERV1	chr18:12765195-12777415	Up	0	3	L540;HDLM2;UH01
SUPHD1.16919.1	2	.	i	MSTD:LTR:ERVL-MaLR	chr18:53752436-53772925	Up	0	3	SUPHD1;MEDB1_;L1236
U2940_U2940_.22831	1	TPM4	s	MLT1A:LTR:ERVL-MaLR	chr19:16185055-16187507	UpEdge	0	3	U2940_;MEDB1_;L1236
L540.15402.3	1	ZNF584; ZNF132	c	L1MC2:LINE:L1	chr19:58914222-58920532	UpEdge	0	3	L540;HDLM2;SUPHD1
L540.13948.2	4	CD70; .	s	HERVE_a-int:LTR:ERV1	chr19:6600108-6601677	EInside	0	3	L540;UH01;SUPHD1
UH01.25120.2	3	.	i	AluY:SINE:Alu	chr2:213801470-213802548	Up	0	3	UH01;SUPHD1;L428
HDLM2.21408.1	1	.; RMDN2	s	L2a:LINE:L2	chr2:38101472-38102560	UpEdge	0	3	HDLM2;SUPHD1;MEDB1_
HDLM2.21719.1	2	SPTBN1	as	MER6A:DNA:TcMar-Tigger	chr2:54891735-54907316	Up	0	3	HDLM2;SUPHD1;L428
SUPHD1.19160.1	5	TMEM18	s	MER33:DNA:hAT-Charlie	chr2:677289-679444	UpEdge	0	3	SUPHD1;MEDB1_;L1236
L540.15528.1	1	.	i	AluY:SINE:Alu	chr2:7604901-7605228	UpEdge	0	3	L540;UH01;L428
HDLM2.22322.1	1	.; GPAT2	s	L1MC4:LINE:L1	chr2:96677639-97754814	Up	0	3	HDLM2;L591;L1236
L591.15585.1	19	KANSL3	s	MIRc:SINE:MIR	chr2:97302660-97303798	Up	0	3	L591;SUPHD1;U2940_
L591.16963.1	1	.	i	L1PB1:LINE:L1	chr20:22896839-22897894	EInside	0	3	L591;L1236;KMH2r
HDLM2.24520.1	1	.	i	L1MA3:LINE:L1	chr20:39607943-39609656	EInside	0	3	HDLM2;L591;MEDB1_

continued on next page…

index_contigID	Exon #	Gene Symbol	TE-Gene Overlap	TE Name	Repeat-Exon Coordinate	Interaction Type	# of Norm.	# of Canc.	Library ID
L591.17211.2	4	.	i	HERVH-int:LTR:ERV1	chr20:43291995-43304515	Up	0	3	L591;U2940_;Karpas1106p_
HDLM2.24686.1	12	BCAS1	s	THE1D:LTR:ERVL-MaLR	chr20:52675395-52675752	EInside	0	3	HDLM2;U2940_;MEDB1_
L540.19000.3	2	NCAM2	s	HAL1-3A_ME:LINE:L1	chr21:22549988-22652972	Up	0	3	L540;HDLM2;L1236
HDLM2.24933.2	1	MAP3K7CL	s	LTR12C:LTR:ERV1	chr21:30448708-30450034	UpEdge	0	3	HDLM2;SUPHD1;U2940_
SUPHD1.25453.1	2	.	i	LTR76:LTR:ERV1	chr3:112020704-112035069	Up	0	3	SUPHD1;U2940_;L1236
SUPHD1.25707.1	11	TPRA1; .	s	MLT1F2:LTR:ERVL-MaLR	chr3:127313643-127314926	Up	0	3	SUPHD1;Karpas1106p_;DEV
HDLM2.26522.2	12	ZFYVE20	s	L2a:LINE:L2	chr3:15138042-15139829	Up	0	3	HDLM2;L591;DEV
U2940_U2940_.34102	1	ARL14	s	MLT1I:LTR:ERVL-MaLR	chr3:160382246-160382636	EInside	0	3	U2940_;MEDB1_;L428
L540.22021.1	2	.	i	MLT1F:LTR:ERVL-MaLR	chr3:180009224-180026116	Up	0	3	L540;HDLM2;SUPHD1
UH01.28858.3	1	GOLGA4	s	MER1A:DNA:hAT-Charlie	chr3:37283188-37285113	UpEdge	0	3	UH01;U2940_;L1236
HDLM2.27364.2	2	.	i	THE1B:LTR:ERVL-MaLR	chr3:72108838-72109939	Up	0	3	HDLM2;UH01;L591
HDLM2.27428.1	1	.	i	AluYc:SINE:Alu	chr3:75463357-75465417	UpEdge	0	3	HDLM2;U2940_;MEDB1_
HDLM2.29775.1	2	.	i	MLT2C2:LTR:ERVL	chr4:117153325-117155076	Up	0	3	HDLM2;UH01;L1236
HDLM2.29935.1	7	ELF2	s	MER65-int:LTR:ERV1	chr4:140088038-140089883	EInside	0	3	HDLM2;L591;KMH2r
U2940_U2940_.36879	4	.	i	AluY:SINE:Alu	chr4:183898709-183899204	UpEdge	0	3	U2940_;L428;Karpas1106p_
L540.22115.1	2	ZNF721	s	AluSc:SINE:Alu	chr4:490916-492812	UpEdge	0	3	L540;U2940_;L1236
L540.24552.1	8	CSF1R; HMGXB3	c	THE1B:LTR:ERVL-MaLR	chr5:149471020-149472372	Up	0	3	L540;MEDB1_;KMH2r
L540.24614.1	2	MED7	s	MIR3:SINE:MIR	chr5:156569408-156569574	EInside	0	3	L540;MEDB1_;L1236
SUPHD1.28169.2	1	.	i	MLT1F:LTR:ERVL-MaLR	chr5:40240108-40240611	EInside	0	3	SUPHD1;U2940_;L1236
UH01.33708.1	1	.	i	LTR12C:LTR:ERV1	chr5:95186686-95188418	UpEdge	0	3	UH01;L1236;L428
L540.26354.1	1	.	i	L3:LINE:CR1	chr6:137978932-137979859	EInside	0	3	L540;L1236;L428
L591.24180.2	1	.	i	MamRep605:Unknown:Unknown	chr6:138051366-138114044	Up	0	3	L591;MEDB1_;L428
L540.26383.2	1	.	i	MER61A:LTR:ERV1	chr6:141167094-141219679	Up	0	3	L540;L591;KMH2r
UH01.35404.1	1	HIST1H4H	as	MSTA:LTR:ERVL-MaLR	chr6:26277174-26286281	UpEdge	0	3	UH01;MEDB1_;L1236
UH01.35833.2	2	ETV7	s	LTR12C:LTR:ERV1	chr6:36330343-36332572	UpEdge	0	3	UH01;SUPHD1;L428
HDLM2.33682.3	1	.	i	LTR27:LTR:ERV1	chr6:78183646-78184368	UpEdge	0	3	HDLM2;L591;SUPHD1
HDLM2.33761.1	1	.	i	LTR12F:LTR:ERV1	chr6:86682224-86685615	Up	0	3	HDLM2;UH01;KMH2r
L540.27718.3	2	RINT1; .	as	HAL1:LINE:L1	chr7:105171944-105172117	UpEdge	0	3	L540;L591;MEDB1_
L540.28031.3	1	.	i	MER51A:LTR:ERV1	chr7:135665356-135666005	EInside	0	3	L540;UH01;SUPHD1
L540.26664.1	1	.	i	MER57-int:LTR:ERV1	chr7:145577-148890	UpEdge	0	3	L540;HDLM2;UH01
L540.29637.1	2	ZBED6CL; LRRC61; ACTR3	c	MER41B:LTR:ERV1	chr7:150019300-150023099	Up	0	3	L540;UH01;SUPHD1
U2940_U2940_.42701	1	.	i	MER4-int:LTR:ERV1	chr7:30602699-30608956	UpEdge	0	3	U2940_;L1236;DEV
UH01.38230.2	1	.	i	L2c:LINE:L2	chr7:50241259-50243769	Up	0	3	UH01;U2940_;MEDB1_
UH01.37581.1	2	.	i	L1MC5:LINE:L1	chr7:5191373-5856594	Up	0	3	UH01;Karpas1106p_;DEV
L591.25504.2	2	GATS; STAG3; PVRIG; SP	c	MER57E1:LTR:ERV1	chr7:99806535-99808808	Up	0	3	L591;Karpas1106p_;DEV
L540.29360.1	2	.	i	ERV3-16A3_LTR:LTR:ERVL	chr8:134676945-134696498	Up	0	3	L540;L591;L428
UH01.40701.2	5	MCMDC2	as	AluY:SINE:Alu	chr8:67838433-67838943	Up	0	3	UH01;MEDB1_;L1236
HDLM2.36852.1	1	.	i	HERVE-int:LTR:ERV1	chr8:6926194-6930731	EInside	0	3	HDLM2;UH01;SUPHD1
L540.29672.1	1	SLC24A2	as	MLT1C:LTR:ERVL-MaLR	chr9:19664535-19665174	UpEdge	0	3	L540;HDLM2;SUPHD1
UH01.41870.1	2	.	i	MSTA:LTR:ERVL-MaLR	chr9:30914798-30925477	Up	0	3	UH01;L591;L1236
UH01.44468.1	1	.	u	MER21B:LTR:ERVL	chrX:103185699-103346273	Up	0	3	UH01;SUPHD1;KMH2r
L540.31047.2	5	ZNF41	s	MER92B:LTR:ERV1	chrX:47341858-47345262	UpEdge	0	3	L540;HDLM2;UH01
MEDB1_MEDB1_.459	82	HUWE1	s	MLT1L:LTR:ERVL-MaLR	chrX:53707002-53743795	Up	0	3	MEDB1_;L1236;L428

Supplementary Table 4.2: Hodgkin Lymphoma Recurrent and Specific TE-Initiated Transcripts

Text 9: Supplementary Table 4.2 Continued

Simplified *LIONS output of Hodgkin Lymphoma cell line RNA-seq (n = 9) and Primary Mediastinal Large B-cell Lymphoma (n = 3) (see Supplementary Table 2.1) which are recurrent to >=3 libraries and specific (absent from B-cell controls, n = 9). Data was grouped by each unique TE that was found to initiate transcription. The TE-intiated contigs from each library were intersected to the UCSC gene annotation for protein coding genes the intersection overlap between the gene and contig was classified as sense (s), anti-sense (as), intergenic (no gene overlap, I) or complex interaction with multiple genes (c).*

A) 5' RACE RT-PCR	
Primer Name	**Sequence**
RLM-outer	GCTGATGGCGATGAATGAACACTG
RLM-inner	CGCGGATCCGAACACTGCGTTTGCTGGCTTTGATG
IRF5-ex_-R1	GATGGTGTTATCTCCGTCCTG
IRF5-ex_-R2	CTCCAGGGGATGCAGAATAA
B) Full-length RT-PCR	
Primer Name	**Sequence**
IRF5-LTR-F	GTCTTCCCTGGCAATACTCG
IRF5-ex8-R	TCTTCCCCAAAGCAGAAGAA
C) Promoter panel/Splicing verification RT-PCR	
IRF5-LTR-F	GTCTTCCCTGGCAATACTCG
IRF5-L2-F	GAAAACGGTTCAGAACCACAG
IRF5-Native-F	CAGGCGCACCGCAGACAG
IRF5-ex2-R	CTCCAGGGGATGCAGAATAA
D) Promoter Contribution qRT-PCR	
Primer Name	**Sequence**
same as promoter panel primers	
IRF5-ex2-F	CAGGTGAACAGCTGCCAGTA
IRF5-ex3-R	TCGTAGATGAGGCGGAAGTC
B-actin-F	AAGGAGATCACTGCCCTGGC
B-actin-R	CCACATCTGCTGGAAGGTGG
E) Genomic DNA bisulfite sequencing PCR	
Primer Name	**Sequence**
IRF5-LTR/L2-Bis-F	ATAGGAGGGAGGTTTTTGAGTAAGT
IRF5-LTR/L2-Bis-R	AAATCCTCTAATCACTCTATACCTTTCTC
IRF5-Native-Bis-F	GAAAGGTATAGAGTGATTAGAGGATTTT
IRF5-Native-Bis-R	CCCAATCTAAACCTAAACTTAAAAACA

Supplementary Table 4.3: Primer List

Name	Type	ENCODE Centre	Data Type	IRF5 Exp.	LOR1A
GM12878	Cell Line	CALTECH ENCODE	RNAseq	+	+
GM12891	Cell Line	"	"	+	+
GM12892	Cell Line	"	"	+	+
H1-hESC	Cell Line	"	"	-	-
HCT-116	Cell Line	"	"	-	-
HeLa	Cell Line	"	"	-	-
HepG2	Cell Line	"	"	-	-
K562	Cell Line	"	"	-	-
LHCN-M2	Cell Line	"	"	-	-
MCF-7	Cell Line	"	"	-	-
HSMM	Primary	"	"	-	-
HUVEC	Primary	"	"	-	-
NHEK	Primary	"	"	-	-
NHLF	Primary	"	"	-	-
GM12878	Cell Line	CSH ENCODE	"	+	+
B cells CD20+	Primary	"	"	+	-
CD34+ Mobiliz	Primary	"	"	+	-
hMNC-PB	Primary	"	"	+	-
Monocytes CD	Primary	"	"	+	-
A549	Cell Line	"	"	-	-
H1-hESC	Cell Line	"	"	-	-
HeLa-S3	Cell Line	"	"	-	-
HepG2	Cell Line	"	"	-	-
K562	Cell Line	"	"	-	-
MCF-7	Cell Line	"	"	-	-
SK-N-SH	Cell Line	"	"	-	-
SK-N-SH RA	Cell Line	"	"	-	-
AG04450	Primary	"	"	-	-
BJ	Primary	"	"	-	-
HAoAF	Primary	"	"	-	-
HAoEC	Primary	"	"	-	-
HCH	Primary	"	"	-	-
HFDPC	Primary	"	"	-	-
HMEC	Primary	"	"	-	-
HMEpC	Primary	"	"	-	-
hMSC-AT	Primary	"	"	-	-
hMSC-BM	Primary	"	"	-	-
hMSC-UC	Primary	"	"	-	-
HOB	Primary	"	"	-	-
HPC-PL	Primary	"	"	-	-
HPIEpC	Primary	"	"	-	-
HSaVEC	Primary	"	"	-	-
HSMM	Primary	"	"	-	-
HUVEC	Primary	"	"	-	-
HVMF	Primary	"	"	-	-
HWP	Primary	"	"	-	-
IMR90	Primary	"	"	-	-
NHDF	Primary	"	"	-	-

Name	Type	ENCODE Centre	Data Type	IRF5 Exp.	LOR1A
NHEM M2	Primary	CSH ENCODE	RNAseq	-	-
NHEM.f M2	Primary	"	"	-	-
NHLF	Primary	"	"	-	-
SkMC	Primary	"	"	-	-
BE2 C	Cell Line	HAIB ENCODE	"	-	-
Jurkat	Cell Line	"	"	-	-
PANC-1	Cell Line	"	"	-	-
PFSK-1	Cell Line	"	"	-	-
SK-N-SH	Cell Line	"	"	-	-
U87	Cell Line	"	"	-	-
GM12878	Cell Line	RIKEN	CAGE	+	+
A549	Cell Line	"	"	+	-
H1-hESC	Cell Line	"	"	+	-
HepG2	Cell Line	"	"	+	-
MCF-7	Cell Line	"	"	+	-
SK-N-SH	Cell Line	"	"	+	-
CD34+ Mobiliz	Primary	"	"	+	-
HMEpC	Primary	"	"	+	-
hMSC-UC	Primary	"	"	+	-
HSaVEC	Primary	"	"	+	-
Monocytes CD	Primary	"	"	+	-
NHEK	Primary	"	"	+	-
HeLa-S3	Cell Line	"	"	-	-
K562	Cell Line	"	"	-	-
SK-N-SH RA	Cell Line	"	"	-	-
B cells CD20+	Primary	"	"	-	-
AG04450	Primary	"	"	-	-
BJ	Primary	"	"	-	-
HAoAF	Primary	"	"	-	-
HAoEC	Primary	"	"	-	-
HCH	Primary	"	"	-	-
HFDPC	Primary	"	"	-	-
hMSC-AT	Primary	"	"	-	-
hMSC-BM	Primary	"	"	-	-
HOB	Primary	"	"	-	-
HPC-PL	Primary	"	"	-	-
HPIEpC	Primary	"	"	-	-
HUVEC	Primary	"	"	-	-
HVMF	Primary	"	"	-	-
HWP	Primary	"	"	-	-
IMR90	Primary	"	"	-	-
NHDF	Primary	"	"	-	-
NHEM M2	Primary	"	"	-	-
NHEM.f M2	Primary	"	"	-	-
Prostate	Primary	"	"	-	-
SkMC	Primary	"	"	-	-

Supplementary Table 4.4: IRF5 Expression and LOR1a-LTR promoter usage

Chromosome	Start	End	Strand
chr2	121804013	121804123	+
chr2	231707185	231707591	+
chr3	70547971	70548278	+
chr4	154563195	154563573	+
chr7	22434412	22434546	+
chr7	128576844	128577151	+
chr7	149847020	149847336	+
chr10	10984958	10985248	+
chr11	23002381	23002680	+
chr11	67796516	67796818	+
chr16	26263803	26263940	+
chr16	27174499	27174780	+
chr19	37627803	37628108	+
chr19	49822524	49822653	+
chr20	4304802	4305130	+
chr1	224970830	224971235	-
chr1	229281547	229281823	-
chr3	167122171	167122500	-
chr4	22929248	22929351	-
chr4	37452697	37452822	-
chr4	153017530	153017844	-
chr5	63821828	63822141	-
chr6	20063149	20063461	-
chr6	160259314	160259445	-
chr7	53959729	53960035	-
chr7	153514112	153514428	-
chr8	129634393	129634799	-
chr10	19696885	19697134	-
chr10	27943768	27943960	-
chr10	121792174	121792590	-
chr16	5586792	5586918	-
chr16	24362440	24362609	-
chr16	50090206	50090516	-
chr20	54677062	54677192	-

Supplementary Table 4.5: LOR1a elements with flanking homology to LOR1a-IRF5

BLAST results in hg19 *of the 69 bp upstream region of the IRF5* associated *LOR1a-LTR (yellow highlight). Each of these matches is located immediately adjacent to a LOR1a element, and are not annotated as being a part of that LOR1a.*

Probe Name	Target Sequence
IRF5_native	TCCCTGGCGCAGCCACGCAGGCGCACCGCAGACAGACCCCTCTGCCATGAACCAGTCCATCCCAGTGGCTCCCACCCCACCCCGCCGCGTGCGGCTGAAG
IRF5_lor1a_a	CCAAGCGAAGAACATTCCATGAGAAGGAACAGGAGACCCCTCTGCCATGAACCAGTCCATCCCAGTGGCTCCCACCCCACCCCGCCGCGTGCGGCTGAAG
IRF5_lor1a_b	TGGCCCGAGGCTCAGCCCGGATCTGCAGTTGCCAGACCCCTCTGCCATGAACCAGTCCATCCCAGTGGCTCCCACCCCACCCCGCCGCGTGCGGCTGAAG
IRF5_total	TGCTGGAGATGTTCTCAGGGGAGCTATCTTGGTCAGCTGATAGTATCCGGCTACAGATCTCAAACCCAGACCTCAAAGACCGCATGGTGGAGCAATTCAA
CSF1R_native	CACCTCACTGGACCCTGTACTCTGATGGCTCCAGCAGCATCCTCAGCACCAACAACGCTACCTTCCAAAACACGGGGACCTATCGCTGCACTGAGCCTGG
CSF1R_the1b	CCTTTGCCTTCCACTATGATTCTGAGGCCTCCTCAGCCATGCTGAACTGTTTACCTGTTCTGGATGTTTCATATAGATGGAGTCGTATGACATTTTGCTA
CSF1R_mirb	CCAGGCCAGAGGGCTGTGGGAGTTCAGAGGTGGACGGACTTTTCAGGCTGAAGCCCAAGTACCAGGTCCGCTGGAAGATCATCGAGAGCTATGAGGGCAA
CSF1R_l3	TATTGAGCACCCACTGTGTTCCAGGCAGTGTGCAGGCCTGACCTCAGGGGGCTCGGAGGCACCCCTGCCTGCTCACTGCTTTGCTTCATGCCTTCCAGGA
CSF1R_total	CTTCACTTCTCCAGCCAAGTAGCCCAGGGCATGGCCTTCCTCGCTTCCAAGAATTGCATCCACCGGGACGTGGCAGCGCGTAACGTGCTGTTGACCAATG
VASH2_native	CCACCCCAAAGGCGCCAAAGGCACCCGGTCCCGGAGCAGCCACGCGCGGCCCGTGAGCCTCGCCACCAGCGGGGGCTCAGAGGAGGAGGACAAAGACGGC
VASH2_MLT2B2	GCACAACAGAGCATGGGACTTCTTGACCTCCATAACCATAACACCATGCTATGTTAGCCGAGCACCATGAGTCCCTGCAGAGAAGGTCATCCTGATTGCC
VASH2_total	GCCGCAGGGCTGAGCTGATGGACAAGCCATTGACTTTTCGGACTCTGAGTGACCTCATCTTTGACTTTGAGGACTCTTACAAGAAATACCTGCACACAGT
FHAD1_native	CTCGGCGGAGGTCGGAGCGTGGGCTTCCTCCTCCCGCCAGGGAAAACAGAGAGGATGAAGGCCTATCTAAAGAGCGCAGAAGGCTTTTTTGTCCTAAATA
FHAD1_mlt1k_a	GACGAAGCTCCATATTTTCTCATTTTCTGCCACGGGAAAAGGAAAACAGAGAGGATGAAGGCCTATCTAAAGAGCGCAGAAGGCTTTTTTGTCCTAAATA
FHAD1_mlt1k_b	TCTGATGACATCACTTGAGCCCTGCAGACTTTTCATTTACGGAAAACAGAGAGGATGAAGGCCTATCTAAAGAGCGCAGAAGGCTTTTTTGTCCTAAATA
FHAD1_total	TTAAAGAACCTCAGAATGGAAAACAATGTCCAGAAAATACTACTGGATGCAAAACCGGATTTGCCAACTCTCTCAAGAATAGAGATCCTAGCGCCTCAGA
CSF1_native	ACTGTAGCCACATGATTGGGAGTGGACACCTGCAGTCTCTGCAGCGGCTGATTGACAGTCAGATGGAGACCTCGTGCCAAATTACATTTGAGTTTGTAGA
CSF1_ltr8	AGCCACTCCATTCTTCTGGAAGCTGCAGGGAAATGGAACCCAGAAACCAGATTGACAGTCAGATGGAGACCTCGTGCCAAATTACATTTGAGTTTGTAGA
ncCSF1_ltr8	GCTGAGATAGTGGCACTTTGCCATAGACTGGTTTCTGCCATAGGCATGTTTAGAAGGACAATGTCCCTCTTCAAGGATGACCTGTTCTACTTTGGGTGAG
CSF1_Total	TGTTCTACAGGTGGAGGCGGCGGAGCCATCAAGAGCCTCAGAGAGCGGATTCTCCCTTGGAGCAACCAGAGGGCAGCCCCCTGACTCAGGATGACAGACA
RALB_native	TGCGGACGGCGGAGGCGGCGGGACTGGTCCCTGCTCTTCAGTGGGTCATCTGTGTGTCACAGCCTCAGAAGACCAGCGAGATGGCTGCCAACAAGAGTAA
RALB_the1c	CTGCCCTGTGAAGTGGTTCCTTCTGCCATGATTCTCTTCAGTGGGTCATCTGTGTGTCACAGCCTCAGAAGACCAGCGAGATGGCTGCCAACAAGAGTAA
RALB_total	TCAATCACAGAACATGAATCCTTTACAGCAACTGCCGAATTCAGGGAACAGATTCTCCGTGTGAAGGCTGAAGAAGATAAAATTCCACTGCTCGTCGTGG
KIRREL3-AS1_msta	GTAAGTTTTCTGAGGCCTCCCCAGCTGTGCCTGCCTACAGAACCCAATTCACCAGGACAGAGGCCTTTCAACTTTCCTCCCTGAGATCTTCCTCCGTGAA
UNC13C_native	TAATGGCATGGTGTGTGCATCTGGAGACCGGAGTCATTACAGTGATTCTCAGCTCTCTTTACATGAGGATCTTTCTCCATGGAAGGAATGGAATCAAGGA
UNC13C_mer73	TGGCTCCCCGTGGCCTCCAGACTTCCCCTCGGGCTCCTGCCGCTCTCTGGACCTCTCTGGGATGTTCGTTCCTCCAAAGATGCCGTGGGTCAGATATCTG
UNC13C_total	AATGACAGTCATTCAGCTACAGAACATAGCAGAAAAGGGAAGCTATGGGGCATGGTATCCTCTTCTGAAAAATATCTCTATGGATGAAACTGGTTTGACT
hlnc1	CAACTCTAGCCACCAGGAGCCAACATTCTTTCAGTGGATAAAAAGGAGTTCCAATACTTTTTCTTACTAAGGAAATGGATTGCAAGGGATGGTGAATTAT
AFAP1-AS1	CTGCCACGTAAGAAGTGTCTTTCGCCTCCCGCCATGATTCTGAGGCCTCCCCAGCCATGTGCAACTGCGTGTTTACTGCTCTGGGCCCAGTGCCTCCCTC
DHRS2	GCAGTGAGACTATTGCCAAGTGGTGAGACCATCACCAAGCGGTGAGACTATCACCTATCGCCAAGTGGCCTGATTCAGCAGGAAGCATCTCAGACACCAA
IL1R2	TAGTGACGCTCATACAAATCAACAGAAAGAGCTTCTGAAGGAAGACTTTAAAGCTGCTTCTGCCACGTGCTGCTGGGTCTCAGTCCTCCACTTCCCGTGT
ZNF281-AS1_mer21b	CGATAGCCTTTGTAATGTCCTTAATAGTAAACCGGAAAACGTGGAGGAAGAAGAGAATCACCACATATCGTATTTAGAGGTCCTGCAGAAAGGGCAGAGC
ZNF281_mer5b	GGGCGTCCCAATGATTTCTACTTCTAAAGAGTGCTAGTGAATGAGGGATTTTGATTGAGGGCTTACTTTGCTGCCTTAGTGTTCTCTCCTAACCAGAAGC
ZNF281_total	TTTTCAAGGACTGATAGATTGTTGAAGCACAGGCGCACATGTGGTGAAGTCATAGTTAAAGGAGCCACTAGTGCAGAACCTGGGTCATCAAACCATACCA
TBP	ACAGTGAATCTTGGTTGTAAACTTGACCTAAAGACCATTGCACTTCGTGCCCGAAACGCCGAATATAATCCCAAGCGGTTTGCTGCGGTAATCATGAGGA
SDHA	TGGAGGGGCAGGCTTGCGAGCTGCATTTGGCCTTTCTGAGGCAGGGTTTAATACAGCATGTGTTACCAAGCTGTTTCCTACCAGGTCACACACTGTTGCA
WBP4	GAGGGTTACCATTACTATTATGATCTTATCTCAGGAGCATCTCAGTGGGAGAAACCTGAAGGATTTCAAGGAGACTTAAAAAAGACAGCAGTGAAGACCG
POLR1B	GGAGAACTCGGCCTTAGAATACTTTGGTGAGATGTTAAAGGCTGCTGGCTACAATTTCTATGGCACCGAGAGGTTATATAGTGGCATCAGTGGGCTAGAA
GUSB	CCGATTTCATGACTGAACAGTCACCGACGAGAGTGCTGGGGAATAAAAAGGGGATCTTCACTCGGCAGAGACAACCAAAAAGTGCAGCGTTCCTTTTGCG
TNFRSF8	GAAACCGCTCAGATGTTTTGGGGAAAGTTGGAGAAGCCGTGGCCTTGCGAGAGGTGGTTACACCAGAACCTGGACATTGGCCAGAAGAAGCTTAAGTGGG

Supplementary Table 4.6: HL-LTR assay target sequences

Probe Name	Accession	Position	ProbeA Tm	ProbeB Tm
IRF5_native	NM_001098629.2	79-178	94	87
IRF5_lor1a_a	IRF5_lor1a_a.1	199-298	83	87
IRF5_lor1a_b	IRF5_lor1a_b.1	302-401	90	87
IRF5_total	NM_001098629.2	1449-1548	86	86
CSF1R_native	NM_005211.3	456-555	86	84
CSF1R_the1b	CSF1R_the1b.1	41-140	84	71
CSF1R_mirb	CSF1R_mirb.1	163-262	92	91
CSF1R_l3	CSF1R_l3.1	133-232	92	90
CSF1R_total	NM_001288705.2	2542-2641	86	87
VASH2_native	NM_001301056.1	461-560	93	93
VASH2_MLT2B2	VASH2_MLT2B2.1	201-300	88	88
VASH2_total	NM_001136474.1	833-932	82	81
FHAD1_native	NM_052929.1	85-184	90	88
FHAD1_mlt1k_a	FHAD1_mlt1k_a.1	457-556	72	88
FHAD1_mlt1k_b	FHAD1_mlt1k_b.1	218-317	74	88
FHAD1_total	NM_052929.1	3685-3784	83	84
CSF1_native	NM_000757.5	526-625	92	76
CSF1_ltr8	CSF1_ltr8.1	404-503	90	76
CSF1_Total	NM_000757.5	1960-2059	89	93
ncCSF1_ltr8	ncCSF1_ltr8.1	6944-7043	83	82
RALB_native	NM_002881.2	111-210	96	77
RALB_the1c	RALB_the1c.1	2-101	76	77
RALB_total	NM_002881.2	470-569	82	81
KIRREL3-AS1_msta	KIRREL3_AS1_msta.1	105-204	86	84
UNC13C_native	NM_001329919.1	2894-2993	83	81
UNC13C_mer73	UNC13C_mer73.1	171-270	92	83
UNC13C_total	NM_001329919.1	7448-7547	84	80
hlnc1	hlnc1_a.1	268-367	82	74
AFAP1-AS1	NR_026892.1	5-104	81	88
DHRS2	NM_182908.4	343-442	90	88
IL1R2	NR_048564.1	95-194	81	88
ZNF281-AS1_mer21b	ZNF281_AS1b.1	243-342	79	85
ZNF281_mer5b	ZNF281_mer5ba.1	439-538	81	80
ZNF281_total	NM_001281293.1	1191-1290	81	85
TBP	NM_001172085.1	588-687	79	82
SDHA	NM_004168.1	231-330	82	80
WBP4	NM_007187.3	516-615	79	83
POLR1B	NM_019014.3	3321-3420	81	80
GUSB	NM_000181.3	1900-1999	84	83
TNFRSF8	NM_152942.2	2031-2130	80	82

Supplementary Table 4.7: NanoString Probes for HL-LTR assay