Human Endogenous Retrovirus Expression in Fetal Tumors and Analysis of Factors Involved in the Post-Transcriptional Regulation of mRNA with Retained Introns

David F. Grabski

Charlottesville, Virginia

Bachelor of Science, Colorado School of Mines 2006

Master's in Education, Harvard University 2014

Doctorate of Medicine, George Washington University 2015

A Dissertation Presented to the Graduate Faculty of the University of Virginia

in Candidacy for the Degree of Doctorate of Philosophy

Department of Molecular Physiology and Biological Physics

University of Virginia

July 2021

Table of Contents

I. Abstract

Human Endogenous Retroviruses (HERVs) are a class of genomic elements that resulted from repeated retroviral infection and integration in the human germline. HERV-K (HML-2) is the most recent proviral integration and has been shown to be biologically active in certain cancers and during fetal development. In my investigation for this thesis, I developed a transcriptional annotation of the approximately 90 Human Endogenous Retrovirus-K (HML-2) proviruses in the human genome. I used this transcriptional annotation to characterize HERV-K expression in fetal tumors and found that mRNA from specific integrated HERV-K proviruses was increased in both Hepatoblastoma and Wilms' tumor. These findings raise the possibility that such expression could be used as targets for immune therapy or biomarkers for each respective disease.

HERV-K, like all retroviruses, expresses mRNA with retained introns. Nucleocytoplasmic export of mRNAs with retained introns requires special mechanisms, since the cell normally restricts the export of incompletely spliced mRNA. Previous investigations from our lab demonstrated that retroviral RNAs contain cis-acting RNA regulatory elements which pair with specific nuclear export proteins to allow efficient export and translation of retroviral mRNA that contain retained introns. In some cases, the export protein is encoded by the retrovirus (HERV-K, HIV, MMTV), while in other cases (MPMV, MLV) the cellular protein Nxf1 is utilized.

Our lab discovered that the RNA element present in MPMV RNA, which facilitates nuclear export, was exapted and duplicated from the cellular Nxf1 gene. This element has been named the Constitutive Transport Element (CTE). In the Nxf1 gene, the function of the RNA element is also to meditate the nucleocytoplasmic export of an Nxf1 isoform with a retained intron. This discovery raised the possibility that other mammalian genes might similarly contain CTE-like elements which facilitate the export of cellular mRNA isoforms with retained introns.

In this thesis, I identified several hundred novel CTEs in mammalian genes, by analyzing sequences that were derived from a retroviral vector trap system, which selected for elements that would allow the export of an mRNA with a retained intron. I then experimentally confirmed that many of these elements function together with Nxf1 to mediate nucleocytoplasmic export and translation of mRNA with a retained intron.

I additionally analyzed how the expression of Nxf1 and the post-transcriptional isoform of WT1(+KTS) alter the expression of cytoplasmic mRNA in a model cell line, using short and long read RNA Seq methods. I found that both proteins increase cytoplasmic mRNAs that contain retained introns from genes that have a direct role in post-transcriptional regulation. Additionally, there was significant overlap between genes with increased intron retention in the cytoplasm following expression of Nxf1 and WT1+KTS and the cellular CTEs discovered in the vector trap.

II. Background and Overall Introduction

Approximately 90% of all mammalian mRNA undergoes alternative splicing [1]. The ability of a single gene to produce multiple mRNAs directly contributes to the complex proteome diversity of mammals compared to organisms such as Drosophila and C. elegans which have a similar number of genes as are present in the human genome [2, 3]. Alternative splicing is now understood to be a highly regulated post-transcriptional process that is crucial in development, cell differentiation, as well as in plasticity during cellular stress [4]. Additionally, the mechanisms that govern the fate of alternatively spliced mRNA isoforms, including nuclear-cytoplasmic trafficking and translation competency appear to be similarly important in defining tissue-specific expression patterns in mammals.

Intron retention (IR), remains one of the least investigated forms of alternative splicing in mammalian systems [5]. Historically, IR was considered a rare event in mammals and was often thought to represent remnants of unspliced pre-mRNA. This is in contrast to IR in both plants and human retroviruses, where IR has long been known to be the principal form of alternative splicing [6, 7]. However, with the recent advancement of deep 2nd generation RNA-sequencing as well as 3rd generation long-read RNA-sequencing, it has become clear that IR is a common form of alternative splicing in mammals as well [8, 9]. Similar to other forms of alternative splicing in mammalian systems, IR has been shown to play critical roles in development and differentiation [10-14]. It appears to play an especially important role in neural differentiation as well as erythrocyte and granulocyte differentiation [10, 15, 16]. It has also been implicated in human diseases including a diverse set of cancers as well as neurodegenerative diseases [17-19].

Our lab has extensive experience studying post-transcriptional regulation of mRNA with retained introns in both viral and mammalian systems. Specific to this thesis, the lab discovered that nuclear export of retroviral mRNA with retained introns is dependent on specific nuclear export proteins and structured, cis-acting regulatory elements in the mRNA [20-22]. In the case of simple retroviruses, the RNA export elements pair with host cell proteins, but the more complex retroviruses utilize specific viral proteins to achieve export of mRNA with a retained intron [21, 23]. For example, Human Endogenous Retrovirus (HML-2) -K, which is the topic of part of this thesis, utilizes a *cis*-acting RNA element, RcRE, and an export protein, Rec, to facilitate this process [24]. HERV-K (HML-2) will hereafter be abbreviated HERV-K in this thesis.

These mechanisms are not unique to viruses. The human nuclear export protein Nxf1 contains a cis-acting RNA element, termed the Constitutive Transport Element (CTE), which pairs with the Nxf1/Nxt1 dimer to facilitate the nucleocytoplasmic export and subsequent translation of an Nxf1 mRNA isoform with a retained intron (intron 10) [25]. The Hammarskjold/Rekosh lab has additionally discovered that multiple RNA binding proteins, including Sam68 and WT1+KTS, appear to increase translation efficiency of mRNA with a retained intron [26, 27].

Intron retention is a dynamic process that is regulated at multiple levels within the cell. For example, it was discovered many years ago that IR mRNA is often restricted from exiting the nucleus in eukaryotes [28-30]. IR mRNAs are also frequently candidates for Nonsense Mediated Decay (NMD), since IR events often generate premature stop codons (PTC) in the mRNA [31, 32]. There is also data that suggests IR may be a cellular mechanism to maintain immediate transcriptome diversity within the cell nucleus, so time considerations with respect to transcription are also important in IR analysis [30]. Given these regulatory mechanisms, there are key experimental principals to IR discovery. Perhaps the most important consideration is isolating RNA from different cellular compartments including total (or nuclear), cytoplasmic and polyribosome compartments. Isolating both total and cytoplasmic RNA from an experimental condition can distinguish RNAs with retained/detained introns in the nucleus and IR RNAs that have been exported to the cytoplasm through CTEs or other mechanisms [5] Additionally, analyzing mRNA in polyribosomes may highlight the IR mRNAs that have the capacity to change the proteome [33]. An additional consideration is looking at IR events (in specific cellular compartments) at different time points with respect to transcription (specifically in controlled transfection/transduction experiments) and cell cycle. Lastly, it remains important to consider cell type specificity in IR analysis.

The most frequently used technology for detecting intron retention is RNA-sequencing (RNA-Seq) [10, 34-36], which is a key experimental approach used throughout this thesis. However, detecting and measuring IR with this technology is more complex than measuring overall levels of gene expression or other forms of alternative splicing. The technical biases that are known to distort gene expression levels such as GC content and amplification biases have an even more drastic effect on IR [37]. In addition, introns frequently overlap highly expressed features in the genome, such as small nucleolar RNAs, SINE RNAs, microRNAs or unannotated exons, which may erroneously inflate count-based measures of intronic expression. Conversely, low complexity regions, common in introns, prevent unique mapping of reads, causing the estimation of intronic expression to artificially drop at these sites. In addition, IR occurs at relatively low frequency in most genes in mammals, and introns tend to be substantially longer than exons, thus requiring a relatively high number of sequencing reads. To adjust for these concerns, specific bioinformatic tools for both IR detection and quantification have been created and are utilized in the ensuing chapters of this thesis [38].

Following RNA extraction for IR based discovery experiments, either poly-A selection techniques or ribosomal depletion techniques are typical to enrich for mRNA. In most circumstances, a positive selection using oligo-dT magnetic bead-based protocol is the preferred option [39, 40]. As polyadenylation usually occurs after splicing (which is largely co-transcriptional, except for retained introns), poly-A selection protocols greatly reduce the concern of unspliced nascent messenger RNA ‘contaminating’ true IR discovery [41, 42]. It is also important to produce stranded libraries [43]. Significant levels of anti-sense transcription, especially long non-coding RNA are often transcribed on the opposite strand of protein coding genes, which confounds IR discovery if un-stranded libraries are used [44]. In the case of Illumina short read sequencing, IR detection generally requires higher numbers of sequencing reads for reliable quantification than other forms of alternative mRNA splicing, since the overall level of IR is often rather low. As always, longer reads and paired end reads are preferred. A low estimate based on a resampling approach suggested that at least 35 million mapped reads are required to detect differential intron usage in a one-versus-one experiment [37]. However, a recent review suggested at least 70 million reads per sample are needed, with ideally more than 150 million reads per sample, in order to avoid intronic alignment biases [45].

Long read sequencing is also rapidly emerging as an important tool for IR discovery and is a key technology utilized in this thesis. The two technologies that are most popular are Pacific Biosciences single molecule real time (SMRT) sequencing and Oxford Nanopore Technology (ONT) sequencing [46]. Third-generation sequencing technologies and specifically direct cDNA and direct RNA sequencing, developed by Oxford Nanopore Technologies, represent a unique opportunity for the detection, characterization and validation of IR. Because these technologies are capable of sequencing individual RNA molecules from start to end, they can elucidate the full structure of transcripts with retained introns. This information can be used to determine whether retained introns disrupt the splicing patterns of neighboring introns and determine start and end sites of IR events. As a result, one can identify open reading frames and better evaluate their susceptibility to NMD. Finally, these technologies can clarify and characterize multiple types of intronic events which display ambiguous coverage patterns in short-read data such as partially spliced long introns due to recursive splicing, introns with low mappability regions and the genomic overlap of IR with other transcriptional events.

In this thesis, I annotate the full transcriptome of HERV-K, including the annotation of individual proviruses capable of producing specific proviral transcripts, with the aid of long-read RNA-seq data. Furthermore, I analyze HERV-K expression in two fetal tumors, hepatoblastoma and Wilms' tumor, using existing short read RNA-seq data. I also describe the discovery of additional mammalian genes that contain cellular CTEs and their potential functional interactions with the nuclear export complex Nxf1/Nxt1. Lastly, I further explore the post-transcriptional role of Nxf1/Nxt1 and WT1+KTS by performing RNA-Seq on cells expressing these proteins from transfected plasmids and analyzing for changes in mRNA expression and transcript isoform expression, including a differential analysis of mRNAs with retained introns in the cytoplasm.

III. Human Endogenous Retrovirus-K (HML-2) Transcriptome Annotation and Proviral Expression in Fetal Tumors

Data from this chapter have been published in the following references:

[1] Grabski DF, Ratan A, Gray LR, Bekiranov S, Rekosh D, Hammarskjold ML, Rasmussen SK. Upregulation of human endogenous retrovirus-K (HML-2) mRNAs in hepatoblastoma: Identification of potential new immunotherapeutic targets and biomarkers. Journal of Pediatric Surgery. 2021.

[2] Grabski DF, Ratan A, Gray LR, Bekiranov S, Rekosh D, Hammarskjold ML, Rasmussen SK. Human Endogenous Retrovirus-K mRNA Expression and Genomic Alignment Data in Hepatoblastoma. Data in Brief. 2020.

A review of HERV-K biology in cancer discussed in this chapter was published in the following reference:

[3] Grabski DF, Hu Y, Sharma M, Rasmussen SK. Close to the Bedside: A Systematic Review of Endogenous Retroviruses and their Impact in Oncology. J Surg Res, 2019.

Abbreviations

Human Endogenous Retrovirus-K (HERV-K), WT (Wilms' Tumor), Hepatoblastoma (HB), Fetal Tumor (FT), Normal Control (NC), Long-Terminal Repeats (LTRs)

Introduction

Hepatoblastoma is the most common pediatric liver malignancy, affecting approximately 500 children in the US each year [47, 48]. Similar to other fetal tumors, hepatoblastoma is thought to arise from embryonic liver progenitor cells that fail to differentiate into hepatocytes [49-51]. As hepatoblastoma precursor cells show different levels of differentiation prior to malignant transformation, this cancer is morphologically complex and histologically subcategorized as one of the following subtypes: fetal, embryonal, or mixed epithelial and mesenchymal undifferentiated small cell [52]. Treatment is multimodal, involving a combination of resection and chemotherapy or transplantation [53]. Five-year survival in North America is between 70-80%, with the best outcomes in early-stage disease [54, 55].

Wilms' tumor (nephroblastoma) is similarly the most common renal malignancy in children [56]. Wilms' tumor is an embryonal tumor thought to develop from undifferentiated nephrogenic rest cells (embryonic renal precursor cells) that subsequently undergo malignant transformation [56, 57]. Evidence that the tumors arise from pluripotent malignant cells includes the typical triphasic histology of Wilms' tumor, which includes a blastemal, an epithelial and a stromal component. Substantial progress has been made in the treatment of Wilms' tumor over the last several decades. Overall survival in High-Income Countries now exceeds 90% for localized disease and 75% for metastatic disease [58]. Current treatment protocols aim to minimize treatment toxicity in low-risk disease and maximize over-all survival in high-risk disease [59].

Though recent progress has been made in treating both fetal tumors, there remains a clear need to identify novel treatment strategies that can offer more children hope for a long-term cure [60, 61]. A full understanding of the molecular drivers of these tumors will be advantageous in the search for new treatments [62]. Additionally, identifying tumor markers which can lead to tumor stratification may lead to tailored treatment protocols and reduce systemic toxicity. This chapter focuses on the expression of Human Endogenous Retrovirus-K (HML-2) (HERV-K) mRNA in both fetal tumors and some of the potential biological and clinical implications of HERV-K expression in these tumors.

Human Endogenous Retroviruses (HERVs) are a class of genomic elements that resulted from repeated retroviral infection and integration in the human germline. It is now estimated that 8% of the human genome is occupied by retroviral DNA, including over 200,000 copies of HERVs from 40 distinct biological subgroups [63-66]. The majority of HERV sequences have accumulated numerous indels making them both replication-incompetent and unable to produce intact viral proteins [67]. Additionally, HERV transcription is often silenced through epigenetic modifications including chromatin remodeling and hypermethylation [68, 69]. However, although no copies have been found to be replication competent in humans, there are several classes of HERVs that still retain Open Reading Frames (ORFs) for viral proteins [70]. Additionally, numerous HERV fragments still contain *cis*-acting signals including promoters, enhancers and RNA transport elements which can affect the human transcriptional and post-transcriptional landscape [71].

HERV transcription and resulting proteins have been shown to have broad effects in both human health and disease. For example, the syncytin protein, which is responsible for cell-cell fusion resulting in syncytiotrophoblasts in the mammalian placenta, is derived from a HERV envelope gene [72]. Also, a HERV promoter activates the amylase gene in the human salivary gland and may have contributed to the expansion of the early human diet to include complex starches [73]. Some studies have also investigated the potential pathophysiological roles of HERV transcription in human disease discovery [74]. Specific investigations have concentrated on neurodegenerative disease [75], cancer [76], autoimmune disease [77] and activation by other viruses [78, 79]. For some diseases like multiple sclerosis, building evidence suggests that expression of HERV Env proteins lead to a pro-inflammatory microenvironment [80, 81] and monoclonal antibody therapy to the HERV protein is now in clinical trials [82]. Similarly, in cancers such as melanoma and breast cancer, growing evidence suggests HERV proteins may represent an effective target for cancer immunotherapy [83-87].

HERVs, like all retroviruses, are RNA viruses that replicate via a DNA intermediate which has integrated into the host genome. Though the majority of retroviral infections occur in somatic cells, when retroviruses infect and integrate into germ-cells, they can be vertically transmitted to progeny in Mendelian fashion [88]. The copy number of the integrated virus in the genome can also be amplified through retrotransposition within a cell. Following the accumulation of nonsense mutations and indels, the viruses become replication-incompetent and fixed or endogenized in the genome. Though numerous indels develop in the fixed provirus, the standard retroviral architecture is largely maintained in many of these integrations including two flanking long-terminal repeats (LTRs), which contain transcriptional promoters and enhancers, as well as the *Gag*, *Pol,* and *Env* regions common to all retroviruses [89].

There have been numerous phylogenetically distinct retroviral integrations in the human genome over evolutionary time, and sub-groups are classified based on the tRNA primer binding site of the HERV reverse transcriptase [90]. The most recent HERV integration in the human genome, HERV-K (HML-2), has a lysine (amino-acid code K) tRNA binding site. HERV-K (HML-2) will hereafter be abbreviated HERV-K in this thesis. HERV-Ks are complex retroviruses with similar structure and sequence similarity to the mouse mammary tumor virus (MMTV) resulting in the Human MMTV-Like (HML-2) distinction [91]. HERV-Ks have integrated into the human genome over the last 5 million years and have accumulated the least amount of mutations amongst the HERV families [92]. Approximately 90-100 copies of HERV-Ks still retain a segment of the proviral genome and many copies still can produce viral proteins [93]. Like the MMTV, HERV-Ks are complex retroviruses, which produce a nuclear export protein Rec [24]. Rec interacts with a *cis*-acting secondary structure in the LTR called the Rec-responsive element (RcRE) to promote the export of mRNA with retained introns. A subset of HERV-Ks contain a 292 bp deletion within the Rec and Env region of the virus, which affects the splicing of the *rec* transcript and produces a separate protein called Np9 [94]. Given the prevalence of this mutation in the HERV-K genome, HERV-Ks have been sub-classified as Type I (those that produce Np9 protein) and Type II (those that still produce Rec) [93]. Though multiple investigations have attempted to define the function of Np9, the function of the protein remains unclear [94, 95].

Of the 90 integrated proviruses, none of the HERV-Ks have been shown to be replication-competent. However, it is clear that when examined in aggregate, all of the viral proteins have the potential to be produced from one or another provirus, and there have been numerous studies demonstrating the production of viral-like particles in specific cell conditions such as during embryogenesis and in different cancer states [96, 97]. Furthermore, several HERV-K copies remain polymorphic in the human population, and new copies are being discovered specifically in centromere and telomere regions of the chromosome as deep secondary and third generation long-read sequencing technology becomes more sophisticated [98, 99].

In addition to the approximately 90 integrated proviruses, there are a ~900 additional HERV-K solo LTRs, which contain both *cis*-acting enhancers and promoters as well as the RcRE [71, 100, 101]. These solo LTRs have been mapped to both intergenic as well as intronic regions of annotated human genes [93]. Previous investigations have found that solo-LTRs are capable of influencing cellular gene expression of a human gene up to 200 kb upstream or downstream from gene loci [100]. Additionally, the ability of the Rec protein to interact *in-trans* with the numerous RcRE elements in both proviruses as well as in solo-LTRs has recently gained interest for its biological implications [102]. The nuclear export axis of HERV-Ks, specifically, Rec and the RcRE, is of particular interest in the Hammarskjold/Rekosh lab, given our focus on retroviral post-transcriptional regulation.

HERV-K mRNA expression has been demonstrated in multiple adult solid organ malignancies [103-105]. Given that HERV-Ks often demonstrate large differences in expression profiles between cancer and non-cancer tissue, as well as their ability to activate the immune system [106-110], attention has recently turned to the potential role of HERV-Ks as targets for immunotherapy [85, 87, 111, 112]. Additional investigations have evaluated the utility of HERV-K expression as tumor markers for disease presence and disease stage [113-115]. For the current understanding of HERV-K biology in cancer, see [89, 116].

HERV-K mRNA expression has also been demonstrated during embryogenesis and is progressively silenced as fetal development continues [96, 117]. Wysocka and colleagues demonstrated that HERV-K Rec transcripts are specifically upregulated during embryogenesis. The authors show through iCLIP experiments that Rec preferentially binds to the RNA from approximately 1,600 human genes and that many of these targets are preferentially translated in embryonic cells [96]. Transcriptional activity of HERV-K proviruses in pediatric tumors are thus of specific interest, as these cancers are thought to arise from embryonic precursor cells that fail to differentiate during organ development. We sought to investigate HERV-K expression in hepatoblastoma and Wilms' tumor to explore the hypothesis that HERV-K mRNA, and possibly Rec, may be more expressed in fetal tumors compared to fully differentiated, non-cancer tissue.

Furthermore, investigation of these elements may lead to the identification of novel tumor markers or new immunotherapeutic targets for the treatment of hepatoblastoma and/or Wilms' tumor.

One of the large limitations in HERV-K research has been the lack of a published HERV-K transcriptome, including information about which proviruses are capable of generating intact retroviral transcripts, such as Rec. This is in part due to the complexity of mapping repeat elements with high sequence homology and the natural complexity of retroviral biology which relies heavily on intron retention, alternative splicing and ribosomal frameshifting to produce proteins. The lack of inclusion of HERV-Ks in the major Gene Transfer Format (GTF) file builds means that HERV-K analysis often remains absent from RNA-seq analysis pipelines. An additional limitation to HERV-K research has been a lack of inclusion of the viral and cellular post-transcriptional regulation in biological analysis. Like all retroviruses, nuclear export and effective translation of HERV-K transcripts remain dependent on the paired Rec and RcRE axis. Thus, it remains critical to not only define HERV-K expression patterns, but specifically what tumors are capable of producing Rec, which could then lead to efficient export and translation of viral proteins.

Given these limitations, the first step in my research was to define the full transcriptome of the approximately 90 previously identified HERV-Ks in the human genome, with specific concentration on which retroviruses could potentially produce fully spliced Rec transcripts. We then explored HERV-K expression in Hepatoblastoma and Wilms' tumor using existing RNA-seq data.

Results

HERV-K Transcriptome Annotation

HERV-Kcon Annotation

As mentioned previously, no exogenous or completely replication competent endogenous HERV-K virus has been found in humans. However, the most recently integrated and most fully intact provirus is HERV-K 113, which has intact open reading frames for all major viral protein regions (Gag, Pro, Pol, Env and Rec). Bieniasz and colleagues utilized the HERV-K 113 provirus along with 10 other recently integrated and human specific, full length HERV-K proviruses (HERV-K101, HERV-K102, HERV-K104, HERV-K107, HERV-K108, HERV-K109, HERV-K115, HERV-K11p22, and HERV-K12q13; nomenclature based on publication) to create a HERV-K consensus provirus copy capable of replication and exogenous infection [118]. We utilized this HERV-Kcon annotation as our baseline 'consensus' HERV-K annotation file. Utilizing the bioinformatics software program Geneious (Biomatters, Auckland, New Zealand), we recapitulated the HERV-Kcon sequence and its annotation. We individually confirmed and annotated the open reading frames of each protein region as necessary by detailed comparison to publications describing the experimentally verified coding sequence of HERV-K proteins [68, 119, 120]. Additionally, we annotated the 292 bp deletion at the beginning of Env that defines Type 1

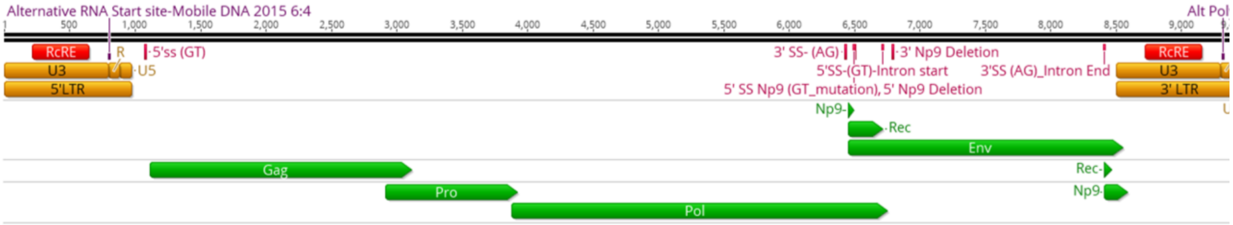

Figure 1: Annotated HERV-Kcon provirus. HERV-Kcon represents a reconstructed HERV-K provirus based on HERV-K113. Key elements include the annotated 5' and 3' LTRs in orange and the RcRE in red. Protein coding regions, specifically the Open Reading Frames for Gag, Pro, Pol, Env, Rec and Np9 are represented in green. The 5' and 3' splice sites as well as the Np9 deletion is also annotated. The figure is based on the HERV-K annotation in Geneious.

proviruses and separately annotated Np9 in these proviruses. We reviewed the sequence and appropriate reading frame to deduce the splice sites for each HERV-K transcript and annotated the 5’ and 3’ LTRs (Figure 1).

We next experimentally verified that the major 5’ and 3’ splice sites for Env and Rec were appropriately annotated as well as the beginning and end of R region (the transcriptional start and stop site for all retroviruses). To achieve this, we utilized a separately reconstructed HERV-K 113 provirus that was gifted to our laboratory. Bannert and colleagues cloned HERV-K113 into a pBSK vector (termed oriHERV-K113) [121]. They used post-insertional mutagenesis to correct specific mutations in the HERV-K 113 provirus to allow the expression of all HERV-K proteins in vitro. We transfected oriHERVK-113 into 293T cells. At 48 hours, we isolated cytoplasmic RNA from the transfected cells and subjected the isolated mRNA to poly-A selection. We used Oxford Nanopore Technology (ONT) Long Read Sequencing (direct cDNA library preparation kit, Single Flow Cell Sequencer) to sequence the poly-A plus mRNA. The long reads were cleaned using the ONT Pinfish pipeline for long read processing including sub-selecting full length reads using the Pychopper algorithm [122]. The sub-selected full-length transcripts were then mapped to the oriHERVK-113 plasmid using Minimap2 (using map-ONT mode) [123].

The full length, mapped reads were imported into Geneious and aligned to the HERV-Kcon annotation for visualization (Figure 2, Panel A). In the figure, the small black lines below the HERV-K annotation represent individual long reads that aligned to the provirus. Despite our bioinformatic selection of full length reads and use of the long-read technology platform, the data does demonstrate multiple smaller (~300 bp) RNA reads that align throughout the proviruses and represent RNA fragments from the various HERV-K transcripts. However, the data also capture several examples of complete, or near complete HERV-K transcripts. More importantly from the perspective of the HERV-K annotation, the data demonstrate that our annotation correctly defined the 5’ and 3’ Splice Site of Env, and both coding exons of Rec (Figure 2, Panel A). This can be seen by the red lines overlaid upon the long-read data, which demonstrate the correct splice junctions from ori-HERV-K113. Figure 2, Panel B is the same long read data, but zoomed in on the 3’ end of the provirus to again highlight the splice reads of Env and Rec. Though not our main aim, the data also nicely confirmed the 3’ end of R (the termination of transcription for retroviruses).

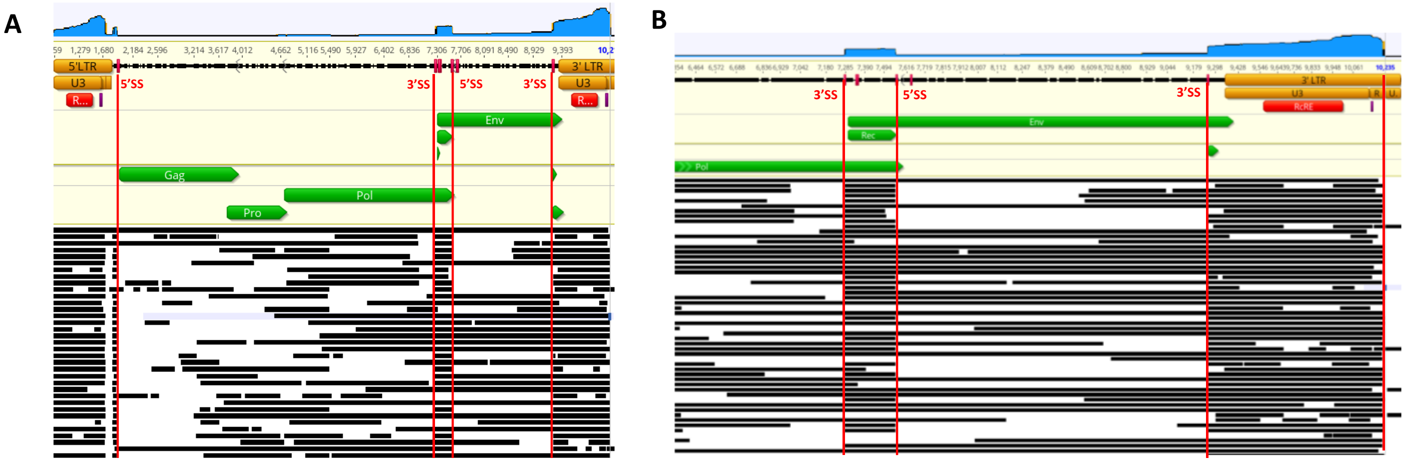

Figure 2: Transfection with oriHERV-K113 demonstrates HERV-K annotation has appropriate transcriptional architecture. 293T Cells were transfected with 15 ug of oriHERV-K113. Cells were harvested 48 hours after transfection. Cytoplasmic RNA was isolated and polyA selected. The cytoplasmic mRNA was used to create a direct cDNA library which was sequenced on an Oxford Nanopore Technology Flow cell to generate 3rd generation long read sequences. Reads were cleaned using Pinfish pipeline and mapped to oriHERV-K113 genome using Minimap2. Aligned reads were imported into Geneious and aligned to the annotated HERV-Kcon sequence. The annotated HERV-K provirus is in the middle of the diagram. A subset of the individual processed reads are represented by the black lines at the bottom of each figure. The key splice sites from the annotation are denoted in the vertical red lines in the diagram. Panel A represents the complete reads across the full provirus. Panel B represents the same data but the 3’ end of the HERV-K provirus to better demonstrate the Rec splice sites.

Transfer of HERV-Kcon Annotation to 91 HERV-K proviruses identified in NCBI

We next imported the individual genomic HERV-K sequences (FASTA files) of the 91 HERV-K proviruses deposited in NCBI (GenBank ID JN675007-JN675097) into Geneious [93]. We performed a pairwise alignment between HERV-Kcon and each individual HERV-K sequence utilizing the alignment tool MUSCLE with a maximum iteration of 10 [124]. Following the pair-wise alignment, we then transferred the annotation from HERV-Kcon to the individual HERV-K provirus. This resulted in a sequence specific definition of the flanking LTRs, Gag, Pol, Env and Rec or Np9 sequence for each respective HERV-K provirus. Lastly, we then manually defined each HERV-K transcript for individual HERV-K proviruses, ultimately creating a FASTA file for each transcript from each provirus in a standard cDNA format used in Ensembl. An example of each FASTA transcript sequence for each individual provirus is demonstrated in Figure 3. Of note, we did annotate the viral regions 5’ U3 and 3’ U5. Though these are not true retroviral transcripts- reads aligning to these regions in an integrated proviruses could suggest read-in transcription from a cellular promoter or transcription that persists through the R termination sequence.

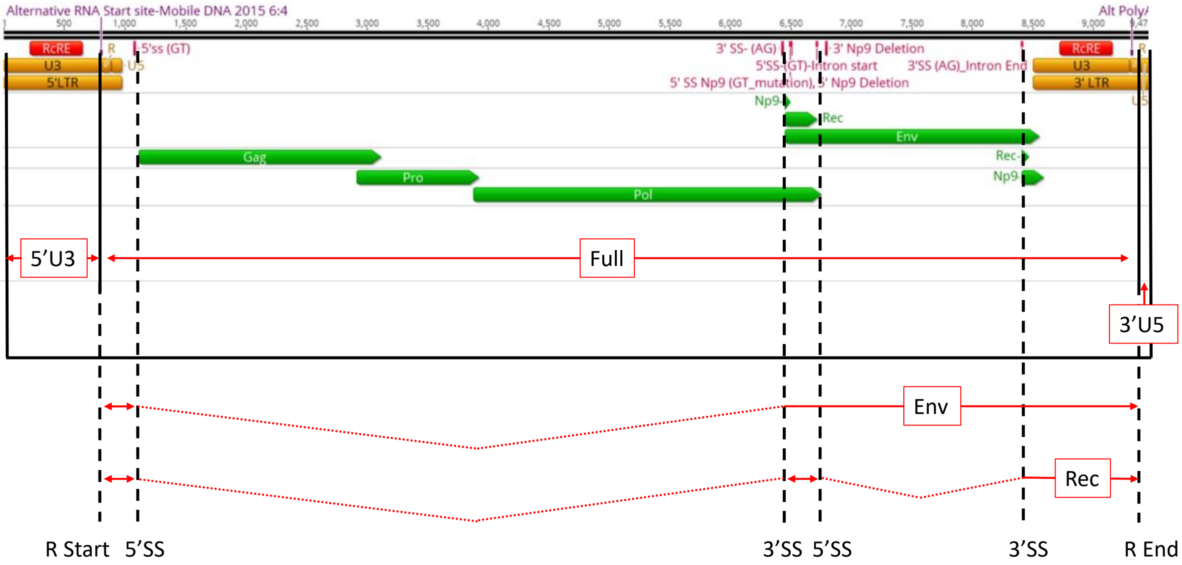

Figure 3: Annotated HERV-Kcon provirus with graphical description of proviral transcripts utilized in HERV-K annotation. The Annotated HERV-Kcon figure is represented as previously shown. The specific transcripts that are annotated for each of the 91 HERV-K proviruses in the cDNA annotation file are demonstrated by the red box overlying HERV-Kcon as a graphical representation of the transcript. Specifically, there is a 5’U3, a Full, an Env, a Rec (or Np9), and a 3’U5 transcript.

HERV-K RNA-seq analysis

To allow appropriate normalization of RNA-seq reads, we concatenated our HERV-K annotation file onto the GRCh38.95 cDNA FASTA file downloaded from Ensembl. This allowed us to further perform appropriate differential expression analysis of HERV-K across different experimental conditions and clinical samples. Of note, though the data is not included in this thesis, analysis of short read RNA-seq data demonstrated that multi-mapping remained an issue for distinguishing individual HERV-K transcripts given the significant sequence overlap for individual transcripts in addition to the general sequence homology across different proviruses. As such, for screening purposes, we first created a simplified HERV-K annotation tool, where we defined the entire individual provirus as an individual transcript and concatenated this FASTA file to the GRCH38.95 cDNA file from Ensembl. We utilized this tool for the majority of the ensuing Hepatoblastoma and Wilms' tumor analysis.

However, the annotation also directly allowed us to identifying all HERV-K proviruses that could potentially produce a full Rec transcript (presence of intact splice sites and 2 coding exons) as well as the additional proviral transcripts. Given our specific interest in post-transcriptional regulation, as well as the preliminary data the Rec protein in produce during embryogenesis [96], we additionally created a similar FASTA file of the human GRCH38.95 cDNA file from Ensembl with the full Rec transcriptome in order to screen tumors for reads that aligned within the Rec region of the HERV-K genome. This was again competed given the limitations of short read RNA-seq data and has the important stipulation that this pipeline does not uniquely define the presence of Rec transcription as there are multiple transcripts that also traverse the Rec region of HERV-K including Env and Pol. One of the major on-going projects in the Hammarskjold/Rekosh lab is the use of the full HERV-K annotation to evaluate long read data and accurately define the complete transcriptome of HERV-K proviruses in various experimental conditions in a single computational step.

HERV-K expression in two Fetal Tumors, Hepatoblastoma and Wilms' Tumor

HERV-K Expression in Hepatoblastoma Compared to Normal Liver Controls

A hepatoblastoma RNA-seq dataset was downloaded from the NCBI biorepository and included 10 hepatoblastoma samples (HB) from children with unresectable disease undergoing liver transplantation (GEO Accession ID: GSE89775). These samples represent aggressive hepatoblastomas that were not amenable to up-front resection, and all children received neo-adjuvant chemotherapy prior to surgery [125]. The dataset also included 3 normal liver controls (NC) from orthotopic adult livers prior to transplant. Utilizing our HERV-K annotation file, HB and NC RNA-seq reads were pseudoaligned using Salmon and samples were normalized using DESeq2. We found that the HERV-K RNA expression profile varied greatly across the hepatoblastoma (HB) samples and normal liver controls (NC). In HB, the median HERV-K read counts across all proviral locations for each sample was 342 (interquartile range (IQR) 235, 515). However, 2 samples had greater than 2,000 reads that aligned to HERV-K proviruses, whereas 2 samples showed 150 read counts or less.

The general HERV-K expression profile across all samples is visualized in the HERV-K expression heatmap in Figure 4. Each cell in the heatmap represents the normalized expression Z-score calculated across samples (range -3 to 3) of a specific HERV-K provirus (y-axis) in individual HB or NC samples (x-axis). mRNA from numerous proviral loci were expressed at average levels across all the samples (Z-score = 0), which we confirmed were HERV-Ks expressed at low levels (average of less than 10 reads across a HERV-K provirus in individual samples). These HERV-K loci are represented by light blue in the heatmap. There was heterogeneity in proviral expression in individual samples, including variation among the hepatoblastoma samples (9.5 expressed proviral loci, IQR 9.3). Variation was also observed in the normal liver controls (3 expressed proviral loci, IQR 3), although the total number of expressed HERV-K loci was less than in HB. No proviral loci were common to all samples and many HERV-K loci were expressed in less than 5 total samples.

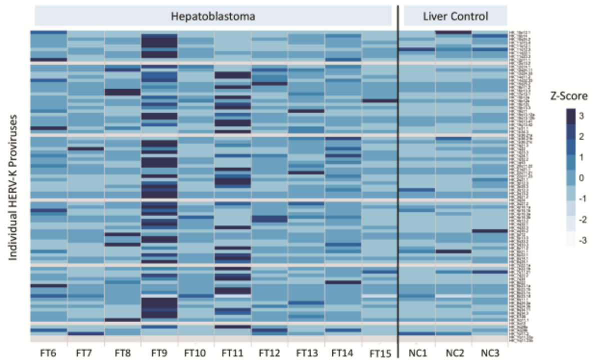

Figure 4: Heatmap of HERV-K expression in hepatoblastoma and normal liver controls. All tumor samples and normal liver controls are represented on the x-axis. All HERV-K proviral loci are represented on the y-axis. Color expression key is located in upper left of figure and is based on the calculated Z score (scale -3 to 3) across all samples in the study. Each individual cell represents the normalized HERV-K proviral expression (x-axis) in the respective sample (y-axis).

We next calculated differential proviral expression across the HB and NC samples. All expressed proviral loci were upregulated in HB compared to NC as demonstrated by the HERV-K expression profile scatter plot (Figure 5). Five proviruses (1q21.3, 3q27.2, 7q22.2, 12q24.33 and 17p13.1) were significantly differentially expressed (p-adjusted value < 0.05, |log2 fold change| > 1.5) across conditions. For these five proviruses, boxplots of the log10 normalized counts for the HB and NC samples across each differentially expressed provirus revealed much higher expression in HB compared to NC samples (Figure 6). For 3 proviruses (17p13.1, 3q27.2 and 7q22.2), the normalized expression across all 3 NC samples was less than 10 reads and was too low for graphical comparison. The absence of expression in NC led to large fold-changes between HB and NC. 17p13.1 showed a 294-fold increase expression in HB as compared to NC (padj = 0.009). This provirus was expressed in all hepatoblastoma samples with one exception and was not expressed in 2 of the 3 normal controls. Similarly, in HB samples, 7q22.2 was expressed 93.1-fold above NC (padj = 0.027), while 3q27.2 was expressed 55-fold above NC (padj = 0.026).

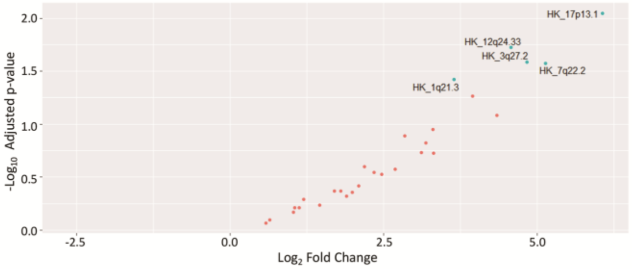

Figure 5: Scatter plot of HERV-K differential expression profile between hepatoblastoma and normal liver controls. Each point on the figure represents the log2 fold change (x-axis) between conditions of an individual HERV-K provirus plotted against the corresponding -log10 p-adjusted value (y-axis). Orange points represent HERV-K proviruses that were not significantly differentially expressed between conditions. Green points represent differentially expressed proviruses (p-adj < 0.05, |log2fold change| > 1.5), which are also labeled by genomic location (HK_1q21.3 represents the HERV-K provirus located at 1q21.3).

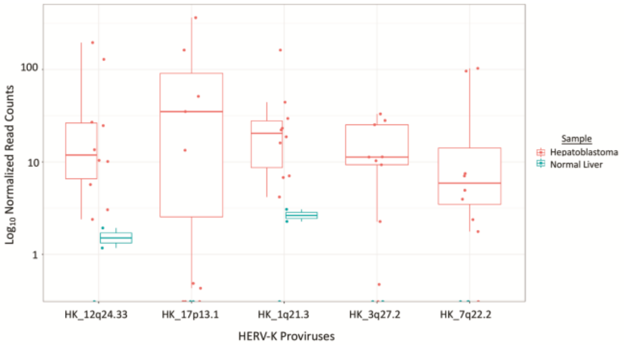

Figure 6: Boxplot and overlaid dotplot of significantly differentially expressed HERV-K proviruses. Individual proviruses are represented on the x-axis. Log10 normalized count values are represented on the y-axis. The normalized count value for each provirus in each sample is represented by an individual-colored point. Hepatoblastoma samples are represented in orange while normal liver controls are represented in green. The central line of the boxplot is determined by the median log10 normalized expression value across all grouped samples (hepatoblastoma or normal liver controls) for each significantly differentially expressed provirus. The 'box' represents the 25th (lower line) and 75th (upper line) percentile of the log10 normalized expression across grouped samples.

Proviral Alignment and Visualization of HERV-K reads in Hepatoblastoma and Normal Controls

Following alignment of all HB and NC samples with a HISAT2-HERV-K index (described in Methods Below) we selected uniquely mapped reads, and imported the unique alignment file (.BAM) into the bioinformatics visualization platform Geneious. This allowed visualization of where the reads aligned across each individual provirus in the different samples. For provirus 17p13.1, the reads aligned across the entire provirus (Figure 7, Panel A). Reads similarly mapped across the length of the provirus at 12q24.33 (Figure 7, Panel B), with a small concentration of reads in the 5’ LTR (long terminal repeat). A similar pattern was seen in provirus 1q21.3, but with a concentration of reads in the 3’ LTR (Figure 7, Panel C). Interestingly, despite 3q27.2 being a relatively complete provirus (9,100 bp), the majority of the reads aligned in the 3’ LTR for all samples (Figure 7, Panel D). Conversely, all of the reads for 7q22.2 aligned uniquely to the 5’ LTR in all samples (Figure 7, Panel E).

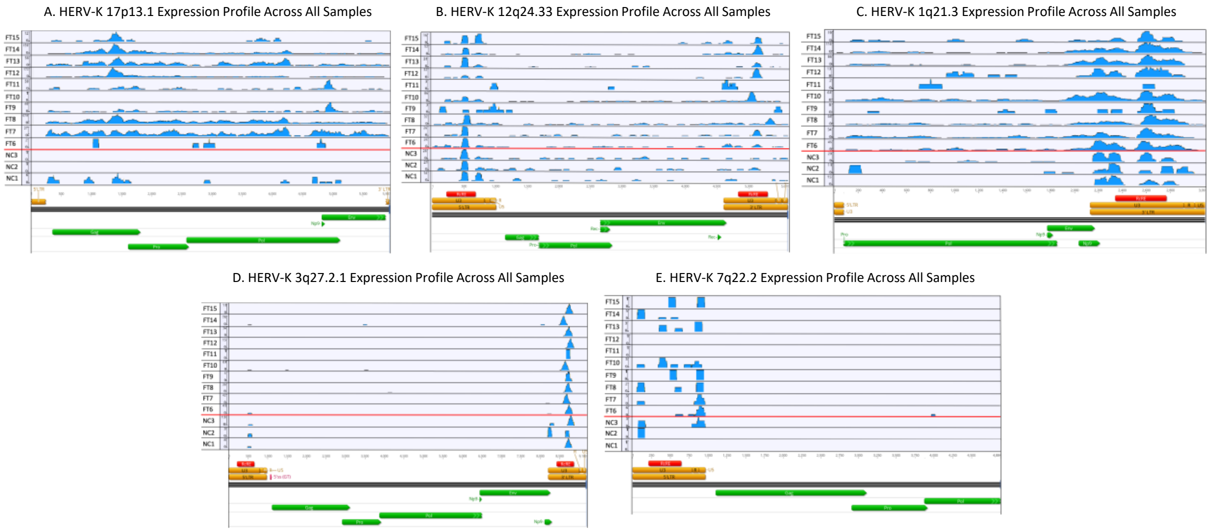

Figure 7: Graphical representation of uniquely aligned reads across HERV-K provirus that were significantly expressed in Hepatoblastoma. (A)17p13.1 (B) 12q24.33 (C) 1q21.3 (D) 3q27.2 and (E) 7q22.2 created in bioinformatics platform Geneious. The x-axis represents the genomic position along the provirus. Major annotated regions of the proviral genome at each provirus are illustrated at the bottom of the panel. Coding regions for viral proteins Gag, Pro, Pol, Env, Rec or Np9 are represented by green bars, but does not necessarily infer an open-reading frame for the protein. Individual reads from each sample are represented on the y-axis. Abbreviations: FT- fetal tumor (hepatoblastoma), NC- normal control (liver).

HERV-K expression correlates with increased immune response in hepatoblastoma

Hepatoblastoma samples FT8 (531 HERV-K reads), FT9 (2,778 HERV-K reads and FT11 (2,503 HERV-K reads) represented the tumor samples with the highest HERV-K expression. Conversely, FT7 (223 HERV-K reads), FT10 (148 HERV-K reads) and FT15 (152 HERV-K reads) represented the tumors with the lowest expression. A differential expression analysis comparing the high to low tumors revealed 775 differential expressed genes. GO Biological Process enrichment analysis of the differentially expressed genes revealed over-representation of cellular processes involved in leukocyte activation and leukocyte mediated immunity (Table 1). Furthermore, global cell processes including immune effector processes and cell activation involved in immune responses were significantly enriched in hepatoblastoma samples that demonstrated increased HERV-K expression patterns.

Table 1: Top 20 Gene Ontology Biological Process Enrichment Analysis following differential gene expression analysis of high HERV-K expressing HB vs low HERV-K expressing HB.

Functional Category	**Differentially Expressed Genes**	**Total Genes in Functional Pathway**	**Enrichment False Discovery Rate (Adjusted p-value)**
Regulated exocytosis	82	901	1.33E-12
Exocytosis	85	1023	2.41E-11
Neutrophil activation	61	594	2.41E-11
Granulocyte activation	61	603	2.55E-11
Myeloid leukocyte activation	69	766	8.96E-11
Myeloid leukocyte mediated immunity	62	647	8.96E-11
Neutrophil mediated immunity	59	591	8.96E-11
Secretion	124	1861	8.96E-11
Myeloid cell activation involved in immune response	61	640	1.60E-10
Neutrophil activation involved in immune response	57	583	3.11E-10
Cell activation	109	1591	4.16E-10
Neutrophil degranulation	56	577	5.59E-10
Secretion by cell	114	1715	6.51E-10
Leukocyte degranulation	59	632	6.51E-10
Vesicle-mediated transport	134	2220	5.32E-09
Immune effector process	96	1392	5.32E-09
Leukocyte mediated immunity	74	965	1.11E-08
Leukocyte activation	96	1416	1.21E-08
Cell activation involved in immune response	66	819	1.48E-08
Leukocyte activation involved in immune response	65	815	3.06E-08

Wilms' Tumor has increased HERV-K expression compared to normal kidney controls and renal cell carcinoma

We gained access to Next Generation Sequencing (NGS) data of 117 Wilms' tumors through the National Cancer Institute- TARGET Initiative [126]. The dataset includes whole genome and transcriptome datasets for clinically aggressive Wilms' tumor (WT). We also gained access to the NCI-The Cancer Genome Atlas (TCGA) dataset and accessed the TCGA-Kidney Renal Clear Cell Carcinoma (KIRC) dataset which includes RNA-seq data for approximately 500 renal cell carcinomas (RCC) and normal kidney (NK) controls (//portal.gdc.cancer.gov/projects/TCGA-KIRC). All datasets included 75 bp paired end RNA-sequencing data. All Wilms' tumor samples were downloaded and a smaller subset of approximately 25 renal cell carcinoma and 25 normal kidney samples were similarly downloaded. All RNA-seq data was trimmed of adapters and quality controlled using Trimmomatic and FastQC.

We first pursued pseudoalignment with Salmon and sample normalization with DESeq2, utilizing our HERV-K annotation file which included the concatenated human cDNA genome and the entire individual HERV-K genome for each provirus (same protocol utilized in hepatoblastoma analysis above). This was done for each experimental condition- WT, NK, RCC. Following sample normalization, the median read counts aligned to all HERV-K loci in WT were 7,676 (IQR 6,619, 8,841). 11 tumor samples had over 10,000 reads aligned to HERV-K loci and none under 5,000 reads. In contrast, the median read counts aligned to NK was 411 (IQR 348, 832). There was slight heterogeneity in the NK samples with several normal kidney samples registering between 3,000-4,000 aligned reads, though none over 5,000. RCC samples had a median HERV-K read count of 445 (IQR 328, 5,861). RCC clearly had significant heterogeneity in total HERV-K read counts across samples, with several tumor samples with over 6,000 reads. In comparison, WT demonstrated high HERV-K expression in all samples and NK demonstrated low expression in all samples.

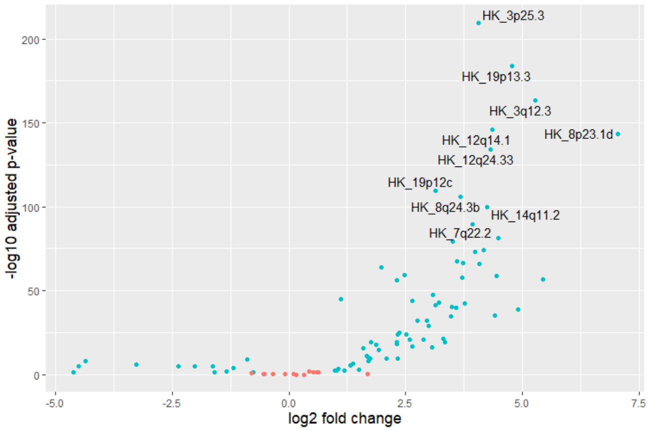

Figure 8: Scatter plot of HERV-K differential expression profile between Wilms' tumor and normal kidney controls. Each point on the figure represents the log2 fold change (x-axis) between conditions of an individual HERV-K provirus plotted against the corresponding -log10 p-adjusted value (y-axis). Orange points represent HERV-K proviruses that were not significantly differentially expressed (p-adj < 0.05, |log2fold change| > 1.5) between conditions. Green points represent differentially expressed provirus. The 10 proviruses that demonstrated the lowest p-adjusted value are also labeled by genomic location.

Following differential expression analysis between WT and NK samples in which we controlled for race and gender- there were 77 (of 91 total) differentially expressed HERV-K proviruses (|fold change| > 1.5, adjusted p-value < 0.05). A scatterplot of differentially expressed HERV-K proviruses between WT and NK is demonstrated in Figure 8. 66 proviruses were upregulated in WT while 11 were down regulated (the down regulated proviruses had mean read counts < 10, suggesting little coverage of these specific proviruses). Several proviruses had very large differential gene expression profiles including the provirus at 8p23.1d (fc 130, p-adj < 0.0001), 8p23.1a (fc 43, p-adj < 0.0001), 3q12.3 (fc 38.8, adj p-value < 0.0001) and 15q25.2 (fc 24, p-adj < 0.0001). Additionally, several proviruses that have potential functional Recs had large differential gene expression profiles, including 8p23.1a and12q14.1 (fc 20.3, p-adj < 0.0001).

This pattern of increased HERV-K expression was largely duplicated when a differential gene expression analysis was performed between WT and RCC samples (again controlling for race and gender). There were 72 differentially expressed HERV-K proviruses between conditions (|fc| > 1.5, p-adj < 0.05); 51 were increased in WT compared to RCC. A scatterplot of differentially expressed genes between WT and RCC is show in Figure 9. The same proviruses demonstrated increased expression in WT compared to RCC as the previous analysis (WT compared to NK). The provirus at 8p23.1d remained elevated (fc 25.2, p-adj < 0.0001), as well as 15q25.2 (fc- 17.6, p-adj < 0.0001) and 8p23.1a (fc 15.2, p-adj < 0.0001).

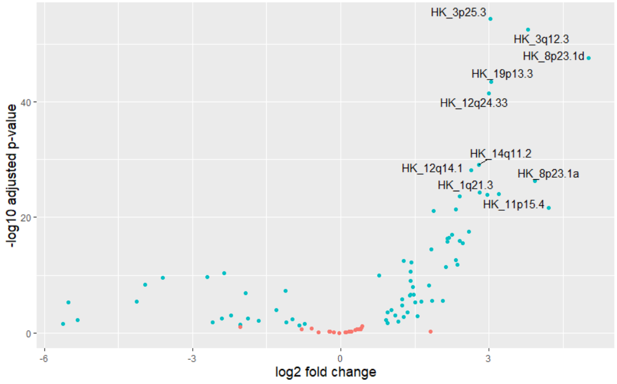

Figure 9: Scatter plot of HERV-K differential expression profile between Wilms' tumor and Renal Cell Carcinoma Samples. Each point on the figure represents the log2 fold change (x-axis) between conditions of an individual HERV-K provirus plotted against the corresponding -log10 p-adjusted value (y-axis). Orange points represent HERV-K proviruses that were not significantly differentially expressed (p-adj < 0.05, |log2fold change| > 1.5) between conditions. Green points represent differentially expressed provirus. The 10 proviruses that demonstrated the lowest p-adjusted value are also labeled by genomic location.

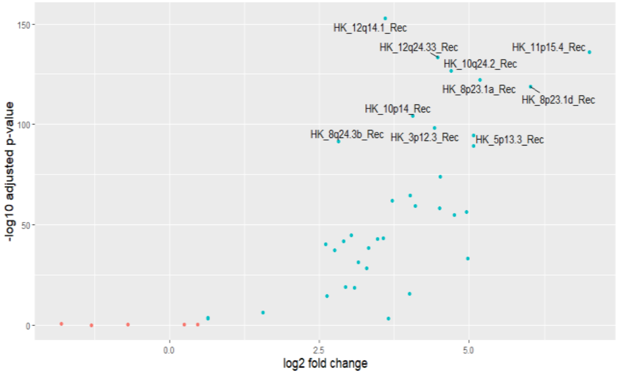

Figure 10: Scatter plot of HERV-K Rec differential expression profile between Wilms' tumor and normal kidney controls. Each point on the figure represents the log2 fold change (x-axis) between conditions of an individual HERV-K provirus plotted against the corresponding -log10 p-adjusted value (y-axis). Orange points represent HERV-K proviruses that were not significantly differentially expressed (p-adj < 0.05, |log2fold change| > 1.5) between conditions. Green points represent differentially expressed provirus. The 10 proviruses that express Rec that demonstrated the lowest p-adjusted value are also labeled by genomic location.

Wilms' Tumor expresses HERV-K reads overlapping the Rec region from multiple proviral loci

One of the striking results of the HERV-K expression profile in WT compared to both NK and RCC was the numerous HERV-Ks that were expressed at high levels that have the potential ability to produce Rec transcripts (intact splice sites and full Rec exons). Given this finding, we created a Rec specific FASTA file and concatenated it to the GRCH38.95 cDNA file from Ensembl (as described in that HERV-K annotation creation above). Following the same pseudoalignment and differential gene expression protocol as utilized above, we analyzed HERV-K reads that overlapped with the Rec region of individual proviruses in all WT and NK samples. Following sample normalization, there were several HERV-K proviruses that were responsible for the majority of reads aligned to the Rec region of individual proviruses. The provirus at 12q14.1 had a median read count of 817 (IQR 698, 937) to the Rec region, while 11p15.4 had 415 reads (IQR 285, 584), 5p13.3 had 360 reads (IQR 283, 464), 10p14 had 320 reads (IQR 274, 404) and 10q24.2 had 235 reads (IQR 204, 289). Following the DGE analysis between WT and NK, these reads that aligned to different Rec regions were clearly increased as demonstrated by the scatterplot in Figure 10. High fold changes were noted between WT and NK, in part due to the relatively high expression values in WT, but also do to the near absent expression of reads aligned to Rec regions in NK samples. Boxplots of the read counts aligned to individual HERV-K Rec loci in WT and NK demonstrate this pattern nicely (Figure 11).

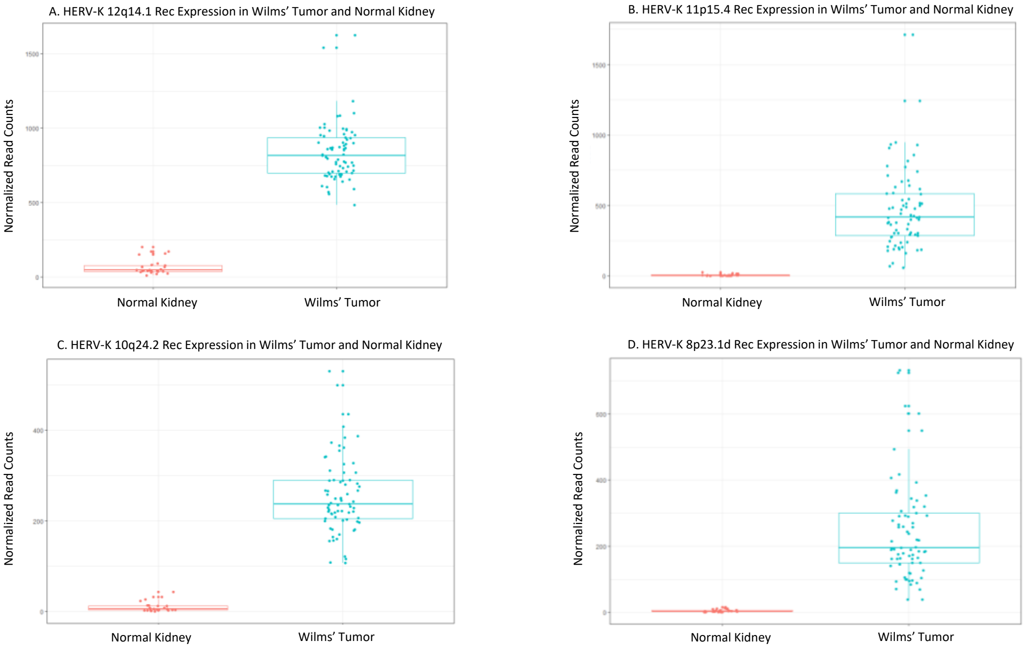

Figure 11: Panel of Boxplots and overlaid dot plots for a subset of significantly differentially expressed HERV-K Rec transcripts reads in Wilms' tumor (green) and Normal Kidney (orange). The x-axis is categorized by sample type. Normalized count values are represented on the y-axis in each panel. The normalized count value for read overlying the Rec proviral transcript in each sample is represented by an individual-colored point. The central line of the boxplot is determined by the median normalized expression value across all grouped samples (Wilms' tumor or normal kidney controls). The 'box' represents the 25th (lower line) and 75th (upper line) percentile of normalized expression across grouped samples. Panel A represents read overlying the Rec proviral transcript at 12q14.1, Panel B represents the Rec proviral transcript at 11p15.4, Panel C represents the Rec proviral transcript at 10q24.2 and Panel D represents the Rec proviral transcript at 8p23.1d. Please note the different scales of the y-axis in each panel.

Discussion

The data in this investigation establish that HERV-K proviruses are expressed in the fetal tumors hepatoblastoma and Wilms' tumor. In hepatoblastoma, the mRNA profile of HERV-K is complex with multiple proviruses transcribed from different loci in tumors from different individuals. The data also show that overall HERV-K expression is increased in hepatoblastoma compared to normal liver controls from several specific proviruses. In Wilms' tumor, a very clear pattern of increased HERV-K provirus expression was noted from 75% of the proviruses compared to both normal kidney controls as well as renal cell carcinoma. Wilms' tumor samples also appear to have reads that overlie the Rec transcripts and further investigations to demonstrate Rec mRNA expression and translation are warranted. The significant increase in HERV-K expression compared to normal controls in both tumors makes HERV-Ks intriguing targets for immunotherapy. In addition, our data suggest that they may also serve as potential biomarkers for disease recurrence or progression, though further studies are required to confirm this. This investigation is the first to demonstrate HERV-K expression in pediatric solid organ malignancies. The data support the hypothesis that fetal tumors may have increased HERV-K expression given the failure of differentiation of these tumor types.

Expression of HERV-K in Hepatoblastoma

Multiple HERV-K proviruses showed increased expression in hepatoblastoma. The HERV-K provirus at 17p13.1 was the most dramatic example of a large differential expression value, with an almost 300-fold change in expression from the provirus compared to normal liver tissue. Similarly, large differential expression values were seen for proviral loci 1p21.3, 3q27.2, 7q22.2 and 12q24.33. The magnitude of the increased expression levels over normal liver controls was prominent in several instances. This is in part because mRNAs from HERV-K proviruses in normal liver control were either not present at all, or present at very low levels (less than 10 read counts across the entire provirus). Our findings of low HERV-K expression in fully differentiated liver is consistent with previous investigations that have examined HERV-K expression in liver tissue [116, 127]. This finding is also consistent with the reported low levels of HERV-K expression in the majority of fully differentiated somatic tissues [116, 127].

Proviral expression profiles also differed across the hepatoblastoma samples. Several hepatoblastoma samples had over 100 read counts aligned to the provirus at 17p13.1, while several other tumors had less than 10 counts (which was more similar to the expression profile in NC). This variation in proviral expression across HB samples was true for total HERV-K expression as well. Two hepatoblastoma samples had over 2,000 normalized read counts summed over all proviruses. In contrast, two HB samples had ~150 normalized read counts across all proviruses with only 1 or 2 proviral locations with over 10 read counts. The HERV-K RNA expression in these tumors was thus more similar to the normal liver controls than to the other tumor samples. In follow-up investigations it will be important to determine whether HERV-K expression correlates with the molecular subtype of hepatoblastoma.

Clinical Applications of HERV-K in Hepatoblastoma

The extremely large fold changes found in HERV-K expression across hepatoblastoma and non-cancer tissue make these genomic elements prime targets as tumor specific antigens, which has been well described in multiple cancers [83, 128-131]. Establishing HERV-K proviruses as a tumor specific antigen in hepatoblastoma leads to intriguing follow up questions, specifically, if these elements may function as novel tumor markers for clinical subtypes of hepatoblastoma or if these elements may act as neoantigens and present novel targets for immunotherapy.

Our current RNA-seq investigation addressed HERV-K expression in tumor verses non-tumor tissue samples and did not address HERV-K expression profiles in peripheral blood samples from patients, limiting our ability to directly address feasibility as a serum tumor marker. Though we were unable to evaluate this in our current investigation, increased HERV-K expression has been demonstrated in both breast cancer tumors as well as the serum of breast cancer patients [132]. In addition, HERV-K expression can effectively differentiate basal cell carcinoma from other breast cancer subtypes [131]. The data in this investigation, combined with studies on other tumors, strongly support further investigation of HERV-K expression in peripheral bloods samples of hepatoblastoma patients to determine if HERV-K expression can be used as an effective tumor marker for disease reoccurrence or treatment resistance.

Regarding HERV-K expression as a potential neoantigen for immunotherapy, a differential gene expression analysis of high HERV-K expressing tumors versus low HERV-K expressing tumors provided promising results. A GO enrichment analysis of the differentially expressed genes demonstrated a strong correlation with cellular pathways involving leukocyte activation as well as neutrophil and leukocyte mediated immunity. Our data suggest that HERV-K mRNA levels may correlate with either direct tumor immunogenicity or an inflammatory microenvironment surrounding the tumors. The possibility that expressed HERV-K proteins in these tumors could be acting as cancer neoantigens capable of activating an immune response is thus an intriguing possibility suggested by our data. It is possible for any of the 5 identified proviruses that are differentially expressed to act as triggers for either the innate or adaptive immune system and produce protein epitopes capable of acting as neoantigens. However, the provirus at 17p13.1, 12q24.33 and 1q21.3 are more likely to represent appropriate targets given that they have expression across viral proteins [133].

HERV-K expression correlating with tumor immunogenicity is supported by recent literature. Rooney et al. analyzed approximately 20 solid organ tumors as well as normal tissue controls from TCGA mRNA-seq datasets for expression of both endogenous retrovirus (ERV) families as well as cytolytic activity. The investigation found that several tumors specific ERVs existed across multiple tumors and that high expression of tumor-specific ERVs significantly correlated with immune activation [134]. In regards to immunotherapy, chimeric antigen receptor

(CAR) T cells that target HERV-K Env proteins have been developed and tested in *in vivo* murine models for both breast cancer and melanoma [85, 87]. In both models, the HERV-K Env CAR T-cells demonstrated tumor specific cytotoxicity, reduced the primary tumor mass and showed reduction of tumor metastases.

One of the limitations of our investigation is that all tumor samples represented advanced disease (unresectable hepatoblastoma require transplantation) and all tumors were exposed to chemotherapy prior to RNA isolation. Little data exists which directly investigates the effect of cytotoxic chemotherapy on HERV-K expression. However, one study in breast cancer suggests that cytotoxic chemotherapy reduced HERV-K expression in peripheral blood samples [132]. Though it is difficult to extrapolate to hepatoblastoma, it may be that neo-adjuvant chemotherapy reduced HERV-K expression in some of our hepatoblastoma samples. A key future investigation will be to evaluate HERV-K expression in early-stage, treatment naïve tumors, which may demonstrate more robust findings.

In our differential expression and gene enrichment analysis we did not find any correlation with cell differentiation or activated cancer pathways to directly suggest a role in embryonal tumorigenesis. Previous literature suggests that HERV-K proviral enhancers can affect transcription of cellular genes up to 100,000 base pairs upstream or downstream of the provirus [70]. Additionally, HERV-K-*env* expression has specifically been linked to perturbations in the TP53 signaling pathways in breast cancer [104]. Though we saw strong expression from the HERV-K provirus at 17p13.1, we did not see alterations in the transcriptional expression of TP53 in our data. It remains unclear, as it does for many HERV investigations, whether HERV-K expression acts as a disease driver in hepatoblastoma or is a result of global epigenetic changes in the cancer cell [135]. What is clear from the current investigation is the tumor specificity and the tumor immunogenicity of HERV-K expression in hepatoblastoma.

Expression of HERV-K in Wilms' Tumor

Compared to the Hepatoblastoma data, which demonstrated a specific set of HERV-K proviruses that were up-regulated compared to normal liver controls, Wilms' tumor samples demonstrated a large increase in expression from the majority of the HERV-K proviruses compared to both normal kidney's and interestingly, renal cell carcinoma. Compared to normal kidney, which had very little baseline HERV-K expression, 67 HERV-K proviruses demonstrated some level of increased expression in Wilms' tumor and 25 demonstrated at least a 10-fold increased expression. From a biological perspective, it was intriguing that several of the highest expressed proviruses in Wilms' tumor were full length proviruses capable of expressing multiple HERV-K proteins including provirus 8p23.1a and 12q14.1 which have open reading frames Gag, Pol, Env and Rec, as well as 3q12.3 which has open reading frames for Gag and Np9.

There have been very few oncologic investigations of HERV-K expression in childhood cancer [136]. However, there have been several studies demonstrating increased expression of HERV-K proviruses during embryonic development in normal tissue [96, 117]. Wysocka and colleagues noted that HERV-K transcript and protein expression were noted as early as the eight-cell stage of embryogenesis up through preimplantation blastocysts. The authors experimental data suggest this proviral expression was a result of DNA hypomethylation at the long terminal repeats (LTRs) of the HERV-K proviruses [96]. Bergallo et al. directly explored HERV transcriptional expression levels (HERV-K, HERV-H and HERV-W) in peripheral blood mononuclear cells (PBMCs) of premature children, newborns, infants and children (4 categories) [117]. HERV-K expression levels directly correlated with younger age including higher expression in premature infants compared to full term newborns. As discussed in the introduction, Wilms' tumor is generally understood to be a failure of differentiation of primordial kidney cells. The results of this investigation again support the hypothesis that HERV-K expression may continue to be increased in cell populations that remain undifferentiated.

Though several investigations have demonstrated HERV-E expression in Renal Cell Carcinoma [137, 138] early investigations of HERV-K in Renal Cell Carcinoma did not demonstrate significant expression patterns [139]. Our data was largely consistent with this finding; the majority of Renal Cell Carcinoma samples demonstrated low level HERV-K expression profiles. Though there was, interestingly, a subset of Renal Cell Carcinoma samples that demonstrated reasonable expression profiles suggesting some heterogeneity in these tumors. When we directly compared expression patterns of HERV-K between Wilms' tumor and renal cell carcinoma we found much higher HERV-K expression in the fetal tumor. Additionally, there was a very similar set of HERV-K proviruses that were again upregulated in this comparison group with respect to the WT and NK comparison. Again, 8p23.1a demonstrated a greater than 15-fold increase which has the ability to encode Gag, Pol, Env and Rec. This is an important comparison as it suggests a specific oncologic profile of the fetal tumor as opposed to simply an oncologic profile of kidney cancer in general.

Potential Rec Expression in Wilms' Tumor

Our data suggests significant reads in the Rec regions of multiple proviruses in Wilms' tumor. Though this is the first investigation to explore possible Rec expression in fetal tumors, several investigations have demonstrated Rec expression in adult malignancies including melanoma, breast cancer and germ cell tumors [113, 140, 141]. Additionally, though there continues to be debate about whether HERV-K expression represents an oncologic significant event or simply a bystander event, several investigations have demonstrated that Rec protein expression may have putative oncogenic properties [142]. Several investigations have demonstrated that Rec binds the human tumor suppressor proteins promyelocytic leukemia zinc-finger protein (PLZF) [143] as well as the PLZF-related testicular zinc-finger protein co-repressor (TZFP) [144] which abolishes the suppression of c-Myc [94]. Additionally, Rec complexes with the androgen receptor (AR) activating the receptor leading to proliferative cell growth [144]. Further investigations in Wilms' tumor to examine c-Myc transcriptional expression targets are warranted by the data in our investigation.

Limitations of Hepatoblastoma and Wilms' Tumor Investigation

There are several limitations in both analyses. The use of non-patient matched liver and kidney control tissue prevented a more thorough analysis of differential expression between tumors and normal organs, given that HERV-K's remain polymorphic in the human population. A lack of aged-matched and patient-matched normal tissue controls is a common issue with current RNA-seq analysis studies of fetal solid organ malignancies and remained true in our analysis. Additionally, as highlighted in the discussion, tumor samples were subjected to neoadjuvant cytotoxic chemotherapy. It is impossible to predict how this exposure directly affected HERV-K expression, though it potentially could have suppressed expression. The small hepatoblastoma dataset did not allow the ability to correlate HERV-K transcription with demographic data including gender and age or clinical outcomes, including stage of disease, reoccurrence, disease resistance and over-all survival. It also limited the potential to correlate HERV-K expression with differentiation/histologic subtypes of hepatoblastoma including fetal, embryonal and undifferentiated disease. While the Wilms' tumor data potentially has the ability to look at these trends, additional analysis is necessary. These will be important future investigations given the preliminary findings in this investigation that HERV-K expression appears to be highly tumor specific. From a bioinformatics perspective, short read RNA-seq data still limits the ability to deal with repeat elements with high sequence homology. In addition, the genomic organization and the overlap of the different genes in the HERV genome makes it impossible to accurately analyze the expression of the individual mRNAs that encoded Gag, Env, Rec and/or Np9 using short read data. However, as long read sequencing continues to become more accurate and available, this analysis is becoming much more feasible.

Conclusion

The current investigation demonstrates that Human Endogenous Retrovirus-K proviruses are transcribed in hepatoblastoma and Wilms' tumor, with increased RNA expression from several proviral loci in the fetal tumors compared to normal organ controls. The large difference in HERV-K expression profiles between hepatoblastoma and normal liver as well as Wilms' tumor and normal kidney make HERV-K proteins intriguing as tumor specific antigen targets. Future investigations are required to explore HERV-K expression as a tool for molecular disease stratification, as well as for targeted immunotherapy in both diseases. Finally, our study highlights the important need to continue to develop tumor banks for pediatric solid organ tumors that include patient matched tissue controls and treatment-naïve tissue samples for appropriate molecular comparison.

Methods

Cell Culture and Transfection of HERV-Kcon

293T cells were seeded on a 10 cm plate at a density of 3x10^6 cells in 10 ml of cell medium (DMEM, 10% Bovine Calf Serum (BCS) and 50 mg/ml gentamicin). 24 hours after seeding, the 293T cells were transfected with 15 ug of the oriHERV-K113 plasmid (plasmid number pRH5847, kindly provided by Bannert and colleagues) using a lipofectamine3000 protocol to the manufacture's specifications (Thermofisher Scientific). Specifically, 15 ug of dsDNA plasmid was added to 30 ul of p3000 reagent which was diluted by 500 ul of Opti-MEM. Concurrently, 23 ul of Lipofectamine3000 reagent was diluted with 500 ul of Opti-MEM. The two solutions were combined and incubated for 15 minutes and the DNA-lipid complex was added to the cell plates. Cells were harvested 48 hours after transfection in preparation for cytoplasmic RNA isolation.

RNA Isolation Protocol and poly-A Selection

Cytoplasmic RNA was isolated from the transfected 293T cells using a Phenol: Chloroform extraction protocol that has been previously described [22]. Key steps to the protocol include washing pelleted cells with ice-cold phosphate-buffered saline (PBS) followed by resuspension in a reticulocyte standard buffer (RSB) (10 mM Tris-HCl (pH 7.4), 10 mM NaCl, 1.5 mM MgCl2) followed by the addition of RSB with 0.1% IgePal in equal volume to perform cell lysis. Cell nuclei were pelleted out using two sequential centrifugations at 13,000 rpm in an Eppendorf centrifuge at +4C. 2 PK buffer (200 mM Tris-HCl (pH 7.5), 25 mM EDTA, 300 mM NaCl, 2% SDS, 400 ug of proteinase K/ml) was added to cell lysates (with nuclei cleared) and incubated at 37C for 30 min. Cytoplasmic RNA was then extracted from the lysate using a phenol and chloroform-isoamyl alcohol (24:1) in a 1:1 volume followed by 2 separate extractions with chloroform-isoamyl alcohol (24:1). RNA was precipitated using 200 proof ethanol (-20C) at a 2.5 volume with the addition of sodium acetate (3M, pH 5.5- final concentration 0.3 M). RNA precipitated in EtOH was stored at -80C. In order to isolate cytoplasmic RNA prior to sequencing preparation- RNA-EtOH slurry was pelleted for 1 hour at 4C at 3,000 RPM (3014g) in a Baxter Cryofuge 6000 centrifuge. RNA pellet was washed twice with 75% ethanol and air dried twice. RNA was resuspended in 10 ul of RNase/DNase free water. RNA concentrations were determined by Qubit RNA HS assay kit (Qubit 3.0 fluorometer, Thermofisher). RNA quality was determined by RNA Tape station. PolyA+ RNA was isolated from the cytoplasmic RNA using poly-T magnetic bead-based protocol according to the manufactures protocol (NEXTFLEX Poly(A) Beads 2.0 Kit, PerkinElmer Inc).

Oxford Nanopore Technology Long Read Library Sequencing and Data Processing

We used a direct cDNA native barcoding ONT protocol (#SQK-DCS109) to prepare our cDNA library from polyA+ cytoplasmic RNA following the manufactures instructions (Oxford Nanopore Technologies, United Kingdom). Following barcode and adapter ligation we loaded the cDNA library on a single SpotON Flowcell with the appropriate amount of sequencing buffer and loading beads. The sequencing was performed with the assistance of the MinIT device (active Base-calling ON) which processed raw signals and converted them to nucleotides (FASTQ format) via MinKNOW v2.2 program. FASTQ reads were processed with the ONT Pinfish pipeline for long read processing which includes sub-selecting full length transcripts using the Pychopper algorithm [122]. The sub-selected full-length transcripts were then mapped directly to the oriHERVK-113 plasmid using Minimap2 (using map-ONT mode) [123]. The full length, mapped reads were imported into Geneious (Biomatters, Auckland, New Zealand) and aligned, using Geneious primary alignment tool, to the HamRek HERV-Kcon annotation for visualization of full length reads.

Hepatoblastoma and Normal Liver RNA-seq Data

The dataset used in this investigation includes RNA sequencing (RNA-seq) data from 10 clinically aggressive hepatoblastoma samples and 3 normal liver controls taken from non-malignant adult liver tissue prior to transplantation. Tumor excision and RNA isolation was performed by the University of Pittsburg Children's Hospital as part of a next-generation sequencing (NGS) study to identify activated cancer pathways in clinically aggressive hepatoblastoma [145]. The raw sequencing data are publicly available and were downloaded from the NCBI biorepository using the NCBI Sequence Read Archive (SRA) Toolkit (GEO accession ID GSE89775). According to the NCBI biorepository, total RNA (1 ug) was isolated from fresh frozen tissue (both hepatoblastoma and normal liver) and sequenced on an Illumina platform to generate 100 base-pair, strand-specific, paired-end reads to a sequencing depth of approximately 40M reads per sample. Prior to analysis in this study, all raw FASTQ files were pre-processed with Trimmomatic to remove adaptors and low-quality reads as well as to assure that only paired-ends reads with a minimum read length of 50 nucleotides were included [146]. The quality of the raw and trimmed reads from each sample was confirmed with the program FASTQC (http://www.bioinformatics.bbsrc.ac.uk/projects/fastqc).

Wilms' Tumor, Renal Cell Carcinoma and Normal Kidney RNA-seq Data

The Wilms' tumor RNA-seq data was accessed from NGS data from 117 Wilms tumor through the National Cancer Institute- TARGET Initiative [126]. Key aspects of the RNA extraction protocol included 2 ug of total RNA isolated from each tumor which were purified, quality controlled, poly-A selected and sequenced on an illumina platform to generate 75 bp paired end reads. Sequence depth was approximately 30M reads per sample. Raw RNA-sequencing data was downloaded from NCBI via dbGaP. Using the same protocol as for the hepatoblastoma data, raw RNA-seq data was pre-processed with Trimmomatic to remove adaptors and poor-quality reads. Quality of each sample was assured with FASTQC. Similarly, we gained access to raw RNA-seq reads for the renal cell carcinoma and normal kidney data from the NCI- TCGA database and the TCGA-KIRC dataset (//portal.gdc.cancer.gov/projects/TCGA-KIRC). The raw tumor sample processing followed the rigorous TCGA requirements. Isolated RNA was sequenced on a HiSeq2000 to produce 75 bp paired end reads. Raw RNA-seq reads were also downloaded through the dbGaP portal similar to the WT data and processed with the same Trimmomatic protocol.

Analysis of HERV-K mRNA expression

We created a HERV-K specific FASTA file using the genomic sequence of the 91 HERV-K proviruses deposited in NCBI (GenBank ID JN675007-JN675097) [93]. To determine the transcriptional profile and differential expression of HERV-K in all samples, we concatenated our working HERV-K FASTA file onto the GRCh38.95 cDNA fasta file downloaded from Ensembl (available at ftp://ftp.ensembl.org/pub/release-95/fasta/homo_sapiens/cdna/). For the short read RNA analysis, we did not annotate the individual potential spliced transcripts that would be expected to be expressed from an integrated provirus (as discussed in the results section), but rather defined the entire proviral sequence as a single transcript to avoid issues with multi-mapping across different transcripts. We then used the pseudo-aligner, Salmon [147] in mapping-based mode with the validateMappings flag to create a count matrix over the full human transcriptome including the concatenated HERV-K file (example code: salmon quant -i GRCh38_HERVK.fa -l A -1 Sample1_1.fq -2 Sample1_2.fq –validateMappings -o Sample1_quant). We then sub-selected read counts assigned to HERV-K loci.

HERV-K Expression Profiles and Differential Expression

Transcript abundance read estimates from Salmon were imported into R (version 3.5.1) using tximport [148]. Transcript abundance estimates were normalized for sample sequencing depth using the R Bioconductor package DESeq2 [149]. This allowed us to determine normalized HERV-K expression across all proviral loci by sample, together with the number of loci responsible for total HERV-K expression and the range of reads across each locus. Differential expression of HERV-K in fetal tumor samples compared to normal organ controls was also analyzed using DESeq2. HERV-K proviruses were considered differentially expressed if the p-adjusted values (calculated using the Benjamini-Hochberg False Discovery Rate implemented in DESeq2) were less than 0.05 and the absolute value of the log2 fold changes were greater than 1.5 [150].

Differential Gene Expression and Gene Ontology Enrichment Analysis in Hepatoblastoma Analysis

Given the apparent heterogeneity in HERV-K expression across different hepatoblastoma samples specifically, we stratified hepatoblastoma samples by overall HERV-K expression (total number of normalized reads across all proviral loci). We selected the top 3 highest HERV-K expressing hepatoblastoma samples and the three lowest HERV-K expressing hepatoblastoma samples. We then performed a differential gene expression analysis again using Salmon and DESeq2 to compare the high expressing to low expressing tumors. Genes with a p-adj value < 0.05 and |log2 fold change| > 1.5 were considered significant and included in a Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) over-representation functional analysis. Gene pathways were considered enriched if they had a p-adj value < 0.05. GO and KEGG analysis was performed using the clusterProfiler package in R [151]. Unless otherwise specified, all plots denoting the RNA expression profile and differential expression were generated using the ggplot2 package in R [152]. The scatterplot was created using the EnchancedVolcano visualization package in R [153].

Proviral Alignment and Visualization

We utilized the HERV-K FASTA file to create a positional index using the alignment program HISAT2 (example code: hisat2-build HB_Data/HERVK_Genome.FASTA HERVK_Genome_tran) [154]. We aligned the HB and NC samples to the HISAT2-HERV-K index to create .BAM files. Uniquely mapped reads were selected with SAMtools (MAPQ Score >= 50). We imported the uniquely aligned .BAM files into the bioinformatics and genomic visualization platform Geneious (Biomatters, Auckland, New Zealand). For the HERV-K proviruses that were differentially expressed between HB and NC, we plotted read distribution of each sample across the respective provirus.

IV. Constitutive Transport Elements in Mammalian Genes and Nxf1 Mediated Intron Retention

Bioinformatic methods utilized in this chapter have been published in the following reference:

[1] Grabski DF, Broseus L, Kumari B, Rekosh D, Hammarskjold ML, Ritchie W. Intron retention and its impact on gene expression and protein diversity: A review and a practical guide. Wiley interdisciplinary reviews RNA. 2021.

Abbreviations

Intron Retention- IR, Constitutive Transport Element- CTE, Non-sense mediated decay- NMD, Human Immunodeficiency Viruses, Rev Response Element- RRE, Oxford Nanopore Technology- ONT, Empty Vector- EV

Introduction

Alternative splicing of messenger RNA (mRNA) is responsible for much of the proteome complexity in mammals [2, 3]. Intron Retention (IR) is a type of alternative splicing in which a complete intron spanning both a 5' splice site and a 3' splice site is maintained in a mature mRNA. Originally described in plants and viruses, IR has now been shown to be a common form of alternative splicing in mammalian systems as well [5, 6, 8, 155]. A recent investigation of over 2,500 human RNA samples demonstrated that IR is ubiquitous and likely to affect over 80% of all protein coding genes [9]. With the introduction of RNA-sequencing technologies, including third generation long read sequencing, the discovery of IR in mammalian systems is increasing.

It is now clear that IR directly contributes to the plasticity of the transcriptome and regulation of gene expression in mammalian systems [8, 13, 34, 156]. Increasing evidence indicates that regulated IR plays key roles during cell development and differentiation, and in response to cellular stress [10-14]. IR is often cell/organ specific and studies indicate that it may play an especially important role in the synaptic plasticity of neuronal cells [10, 15, 16]. IR has also been implicated in many human diseases including cancer and neurodegenerative diseases [17, 18]. In cancer, increased intron retention is responsible for diversifying cancer transcriptomes and has been specifically shown to change the expression of tumor suppressor genes [18, 19].

Despite the prevalence of IR mRNA isoforms in mammals, the regulatory mechanisms governing IR post-transcriptional regulation remain largely unexplored in mammalian systems. It was discovered many years ago that IR mRNA is often restricted from exiting the cell nucleus [28-30]. IR is also often a candidate for non-sense mediated decay (NMD) in the cytoplasm given that IR often have premature stop codons (PTC) [31, 32]. This led to the proposal that IR was chiefly associated with Regulated Unproductive Splicing and Translation (RUST) [157]. However, there are clear examples of protein isoforms translated from IR mRNA in mammals [10, 25, 158-162] showing that mechanisms exist to govern nuclear export and translation of these alternatively spiced isoforms. Though very little is known about these mechanisms in mammalian cells , nuclear export of IR mRNA have been studied in retroviruses for more than 30 years [5].

A key principle of retroviral biology is that an integrated provirus generates a single RNA transcript that is used for both replication as well as the generation of several structural and enzymatic proteins. Some of this full-length mRNA remains unspliced and retains at least one complete intron [5]. Early investigations in Human Immunodeficiency Virus (HIV), found that HIV produces a nuclear export protein Rev that is generated from a fully-spliced mRNA. The Rev protein has a Nuclear Localization Signal (NLS) that traffics the protein into the nucleus where it pairs with a cis-acting secondary structure in the viral IR mRNA, termed the Rev Response Element (RRE) [21, 163, 164]. This paired nuclear export protein and cis-acting RNA regulatory element enables viral IR mRNA to recruit cellular nuclear export complexes and be both exported from the nucleus and be efficiently translated [21]. Further studies, including by our group, demonstrated that either eliminating Rev or mutating the RRE nullified nuclear export of the viral IR-mRNA, confirming that nuclear export of the intron containing mRNA was dependent on this paired interaction.

Further investigations into complex retroviruses demonstrated analogous paired *trans-acting* nuclear export proteins and *cis-acting* RNA secondary structures to overcome nuclear restriction of intron containing mRNA in these viruses. Investigations into Human T-cell leukemia virus (HTLV-1) demonstrated the virus also expresses a nuclear export protein known as Rex (similar to Rev) from a fully spliced mRNA that binds to a Rex Response Element (RxRE) [165].

The same basic mechanism was also described in the Mouse Mammary Tumor Virus (MMTV) (export protein- Rem, RNA structure- RmRE) [166]. Important to this thesis, a similar mechanism was later demonstrated in Human Endogenous Retrovirus-K (HML-2). As described in Chapter 1, HERV-K produce a fully spliced mRNA called Rec, which acts as a nuclear export protein for intron containing HERV-K mRNA which contain the paired RNA-signal (RcRE) [23, 24].

Despite these exciting discoveries, what initially remained unclear was how simple retroviruses, which do not produce regulatory proteins, would be capable of mediating nuclear export of intron containing mRNA. A serendipitous discovery, while investigating HIV nuclear export, noted that when a small fragment of Mason Pfizer Monkey Virus (MPMV) was used for a poly-A and transcription termination signals in HIV, Rev was no longer necessary for nuclear export [20]. Further investigations elucidated that this fragment of the MPMV genome contained a cis-acting RNA element, termed the Constitutive Transport Element, which was shown to be essential for the nucleocytoplasmic export and subsequent translation the unspliced MPMV RNA that retains an intron [20, 167]. Because MPMV is a simple retrovirus and does not express any regulatory proteins, it was clear that the virus must rely completely on cellular proteins to mediate export via this CTE. The mRNA export receptor Nxf1 (previously Tap) was later discovered to dimerize with a co-factor Nxt1 and bind the MPMV CTE [168].

Importantly, this mechanism was later demonstrated to be conserved in mammalian systems as well. It was shown by our lab that the NXF1 gene itself contains a CTE within an internal intron with high primary sequence and secondary structure homology to the MPMV CTE [161]. The Nxf1 CTE facilitates nuclear export and translation of an NXF1 mRNA isoform that retains the CTE-containing intron (intron 10) [25]. The fully spliced NXF1 mRNA isoform expresses the protein that dimerizes with the cofactor Nxt1 and directly binds to a stem loop structure on the CTE to mediate nuclear export of NXF1-IR mRNA [169]. This leads to the expression of a novel short Nxf1 protein (sNxf1), which is expressed in both mouse and human neurons and some other tissues [25]. Interestingly, homologues CTEs are presented in the Nxf1 genes in most mammals and teleost fish (including Zebrafish), demonstrating a highly conserved mechanism for the export of IR mRNA [170].

In addition to nuclear export of IR mRNA, it appears there are several other proteins that facilitate translation of IR mRNA containing CTEs, potentially serving to circumvent NMD. Our group previously demonstrated that the protein SAM68, a STAR (signal transduction and activation of RNA metabolism) family ribosomal binding protein (RBP), increases the cytoplasmic utilization of an HIV Gag-Pol CTE reporter assay. Specifically, RNA analysis demonstrated that the addition of Sam68 only slightly increased the cytoplasmic level of Gag-Pol-CTE RNA levels, compared to a 60-fold increase in Gag-Pol protein levels, suggesting SAM68 was not directly involved in nuclear-export of IR-mRNA, but had a large effect on translation efficiency [26]. Additionally, the WT1+KTS isoform, described in detail in Chapter 3 of this thesis, was also noted to increase translation of IR-mRNA [27].

In summary, it is clear that many different mechanisms are involved in the regulation of mRNA with retained introns and that they can have a number of different fates. These are shown and described in Figure 12. In this chapter, I further explore intron retention in mammalian systems. I re-explore data from a previously completed retroviral vector trap experiment in the Hammarskjold/Rekosh lab that was designed to discover additional mammalian genes that contain cellular CTEs. Additionally, I perform an over-expression experiment of Nxf1 to discover human mRNAs with retained introns that persist in the cytoplasm.

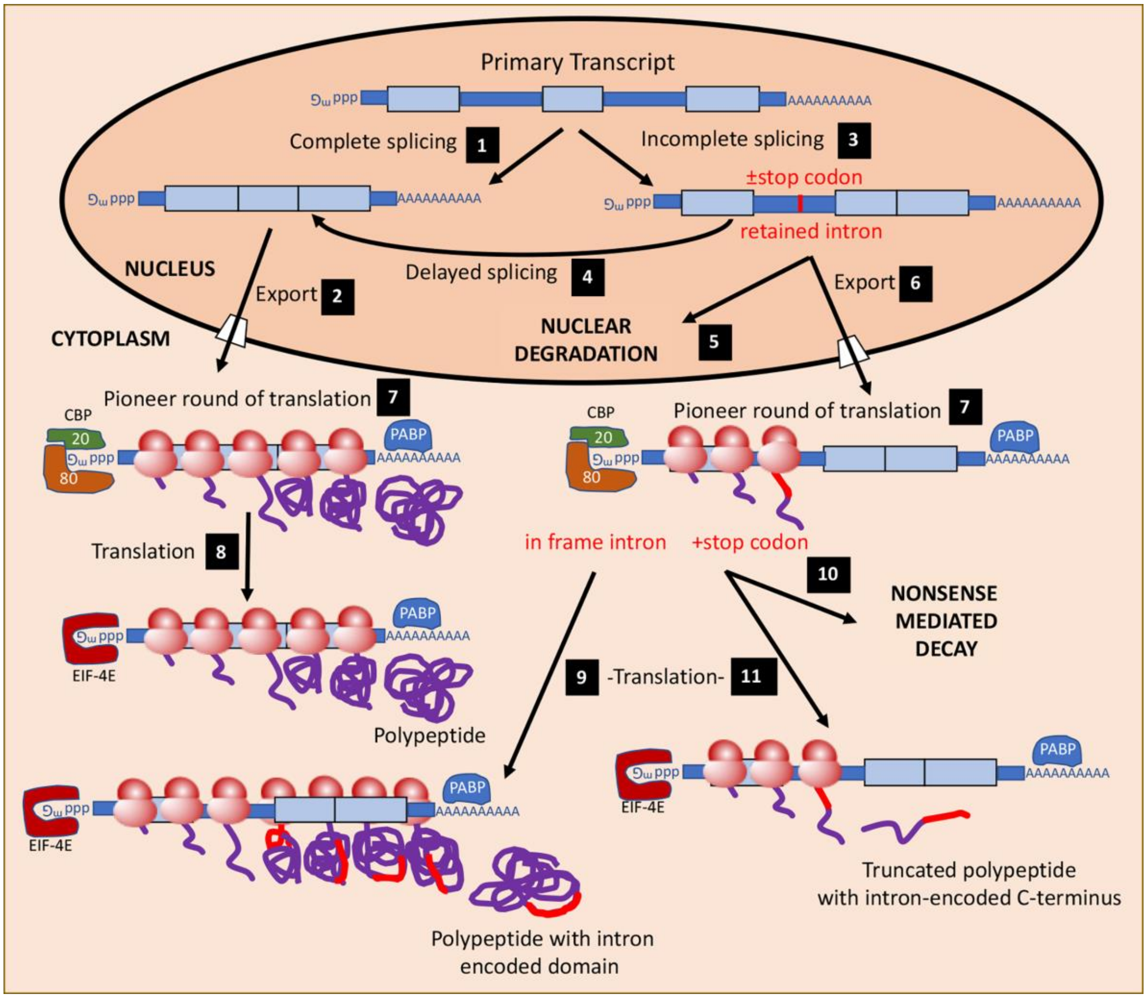

Figure 12: The many fates of mRNA with retained introns. In the nucleus, the primary transcript may either be completely spliced (1) and rapidly be exported to the cytoplasm (2) or it may be incompletely spliced to retain one or more introns (3). In response to a signal, it may then undergo further 'delayed' splicing (4) and export (2) or it may be degraded in the nucleus (5). It may also be exported directly without further splicing (6). In the cytoplasm, all mRNAs undergo a pioneer round of translation (7). The fully spliced mRNA may be translated into a polypeptide (8). If the mRNA retains an intron that is in frame, it may be translated into a polypeptide that contains a novel internal domain encoded by the intron (9). If the mRNA retains an intron that contains an in-frame stop codon, it is often degraded by nonsense mediated decay (NMD) (10). If NMD is avoided it may be translated into a truncated polypeptide with a C-terminal domain encoded by the intron (11). The Cap binding proteins CBP20 and CBP80 that are present on the mRNA during the pioneer round of translation are replaced by EIF-4E as translation proceeds in the cytoplasm. PABP is the poly-A binding protein. Different isoforms of PABP bind poly-A in the nucleus and cytoplasm. Many other proteins that are not shown are also bound to the mRNA and facilitate translation.

Results

Identification of novel Constitutive Transport Elements (CTEs) in mammalian genes using a retroviral vector trap system

Retroviral Vector Trap System

Following the discovery of the Nxf1-CTE, the CTE sequence was used to search mammalian genomes for similar sequences using BLAST. These searches did not find any sequences with significant regions of homology with the MPMV or NXF1-CTE, except in related NXF genes. However, cis-acting RNA regulatory elements often show complex folding, where secondary and tertiary structure determines the protein binding specificity. Thus, primary sequence homology is often minimal, making it hard to find elements bioinformatically. To functionally identify CTEs, our laboratory devised a functional vector trap system to identify cis-acting RNA elements (cellular CTEs or cCTE) in expressed mammalian gene that could substitute for Rev and the RRE. The system is based on a modified HIV vector (pTR167) derived from the NL4-3 HIV provirus [155]. The vector retains part of the HIV Gag and Env genes and has an internal cassette capable of expressing the Hygromycin resistance gene from an internal SV40 early promoter. In transfected cells, the vector produces a full-length viral mRNA from the HIV 5'LTR that is capable of being packaged in viral particles if a packaging system is provided. However, this RNA retains an internal intron that in the original version of pTR167 require a cis-acting secondary structure in the mRNA (HIV-RRE) and the related export protein (HIV-Rev), that binds to this element, for cytoplasmic expression. With Rev co-expression, the vector with the RRE packages in mammalian cells expressing HIV Gag and Pol proteins if VSV-G env, Rev and the transcription factor, Tat, is provided. The packaging efficiency reflects the amount of mRNA that is exported to the cytoplasm and can be measured by using the viral supernatant stock to transduce cells and selecting for hygromycin resistant colonies. Using a COS cell line that stably expresses the HIV Gag and Pol proteins (B4.14), the RRE-containing vector produces around 3x10^3 Colony Forming Units (CFU)/ml of viral supernatant, when Rev is provided (Figure 13). However, when the RRE is removed from the vector backbone and is replaced with a stuffer fragment (derived from the U2 RNA sequence), the viral titer drops to 1-3 CFU/ml (Figure 13). If the stuffer fragment is replaced with the MPMVCTE, the titer of around 3x10^3 CFU/ml is restored. These data demonstrate that nuclear export of the vector RNA genome is appropriately dependent on a paired RNA export signal (either the RRE or the CTE) and its corresponding export protein (Rev or Nxf1:Nxt1(constitutively expressed in COS cells). These results provided proof of principle and suggested that it should be possible to select cellular CTEs using this vector trap.

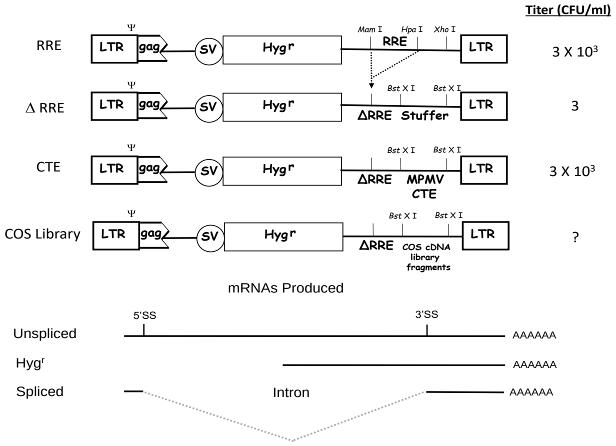

Figure 13: Schematic drawing of HIV vectors used in Vector Trap to discovery cellular CTEs. The retroviral vectors, containing segments of *gag* (DGAG) and envelope (DENV) sequences, are derived from the HIV-1 NL4-3 isolate, and a hygromycin B resistance gene (HYGROMYCIN) driven by the SV40 promoter (SV) is inserted in the middle portion of the vector. Three different kinds of mRNA are produced from the vector (see diagram). The "unspliced" mRNA is retained in the nucleus without the presence of a *cis*-acting element. In the case of the RRE, the export of unspliced message is dependent on the presence of Rev protein. In the presence of the MPMV CTE, the unspliced mRNA is exported to the cytoplasm using cellular factors including Nxf1 and Nxt1. Once the unspliced message reaches the cytoplasm, it is subsequently packaged into virions. The resulting virions confer hygromycin resistance when transduced cells are grown in selection medium. The long terminal repeat (LTR), the packaging signal (Y), the Rev response element (RRE), the stuffer sequence from pCDM8 (stuffer), and the MPMV constitutive transport element (CTE) are indicated. Restriction sites relevant to the construction of each vector are also shown. 5'SS, 5' splice site; 3'SS, 3' splice site.

The effort to use this system to identify other elements capable of functioning as cCTEs in mammalian genomes was started over 20 years ago, by Dr. Yeou-Cherng Bor, a post-doc in the Hammarskjold/Rekosh lab. Random cDNA fragments (either ~500 or ~1,000 bp long) from a COS-cell cDNA library were inserted into the pTR167 vector in place of the stuffer fragment. The HIV vectors with the inserted cDNA library fragments were then transfected in bulk into the HIV B4.14 HIV GagPol cell line together with plasmids expressing Tat and the VSVG Env protein (see Figure 14). After 72 hours, viral stocks were harvested from the transfected cells and used to transduce either Hela cells or 293T cells. If any vector RNA was exported to the cytoplasm and packaged, then the vector conveyed hygromycin resistance to the transduced target cells, as the hygromycin gene is independently driven by an SV40 promoter and its expression is not dependent on the presence of an export element. The target cells underwent hygromycin selection for 14 days. Surviving cell colonies theoretically had one copy of the HIV vector integrated in their genome. To assure that the inserted cDNA fragments indeed enabled export of intron containing mRNA in the context of an integrated provirus, target cells were then exposed to vector mobilization by providing the respective target cells with the necessary packaging system (HIV gag-pol-rev-tat, VSV-G). After 72 hours, viral supernatant was collected and used to transduce either Hela cells or 293T cells for a second round of hygromycin selection (Figure 14).

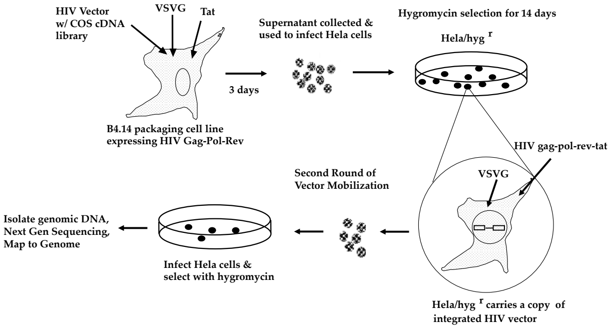

Figure 14: Experimental design for the cloning of cellular CTEs. The B4.14 packaging cell line, which constitutively produces HIV-1 GagPol proteins, was transfected with vectors containing either the RRE, CTE, no RRE/CTE (DRRE) or the vector library containing COS cDNA fragments. Three days post-transfection, supernatants were collected and used to infect Hela cells, which were subsequently selected in hygromycin-containing medium for 14 days. Resistant colonies were then subjected to a second round of vector mobilization by cotransfecting plasmids expressing HIV-1 GagPol and VSVG envelope proteins to rescue the integrated HIV vector sequence from the genome. Supernatants from transfected cells were used to infect unaffected Hela cells that were again selected in hygromycin-containing medium. Genomic DNA was isolated from resistant colonies and used as a template to amplify cDNA fragment using PCR primers flanking the cloning sites in the vector. COS cDNA library fragments, which contained cellular CTEs, were then sequenced on MiSeq and Oxford Nanopore Long Read Sequencing Platforms and mapped to the human genome.

At the time these experiments were performed, NGS sequencing had not been developed and it was a tedious undertaking to grow up each selected colony and determine the sequence fragment present in each individual proviral integration. Additionally, there were large gaps in the human genome database, so many of the sequenced fragments could not be linked to a specific gene, although some potential CTEs were identified and tested further, including genomic fragments present in the SIRT7 and ACTN4 genes [27]. To preserve the selected material, colonies were scraped from several plates and frozen. In addition, individual colonies were also picked and frozen in pools, waiting for the day when technology would allow a more rapid and detailed analysis of the selected elements.

Processing genomic DNA from vector trap colonies

We decided to resurrect this project when I joined the lab and the work described below was conducted by myself in collaboration with a Research Scientist in the lab, Dr. Sarah French. We used both 2nd generation and 3rd generation DNA sequencing technologies to identify potential CTEs in genomic DNA from colonies obtained after 2 rounds of selection. To do this, primers flanking the library fragment insert in the vector were synthesized. The primers were then used to PCR amplify the inserts and the resulting DNA was sequenced on both a MiSeq and Oxford Nanopore Technology (ONT) platform. On the "short" read MiSeq platform, we had the advantage of being able to sequence DNA from multiple different pools in one run, enabling us to get potential cCTE sequences from most of the pooled colonies. ONT long read sequencing was only done on a subset of colonies, but allowed us to sequence across the entire cDNA insert (500-1,000 bp) and get full length sequences of many potential cCTE elements. Following sequencing, we used Cutadapt to trim each sequence to the 5' and 3' BstX1 linker sites [171]. We used the Burrows-Wheeler Aligner (BWA) and BBMap to align the trimmed cDNA fragments to the human genome [172]. We sub-selected cDNA fragment reads that were in the same sense in the identified mammalian gene and the mRNA produced by the HIV vector (5' to 3' in the respective mRNA from the identified gene).

A full list of elements sequenced on the MiSeq and ONT platform are represented in Appendix Table 1. A subset of the genes containing a putative cCTE that aligned in the sense direction on either platform are represented in Table 2. A total of 545 individual mapped genes with potential cCTEs in the sense direction were sequenced on the ONT platform, with an additional 94 mapped genes in the sense direction sequenced on the MiSeq platform. There were 67 elements that overlapped between the two lists leaving a unique group of 572 potential cellular CTEs mapped in genes. Of note, there were a smaller subset of cellular cDNA fragments that were isolated from the vector trap and aligned in the antisense direction (data not shown). This may represent cCTEs in "antisense" genes, but this will require further analysis. Similarly, there were a subset of reads that mapped to unannotated regions (data not shown), which require additional investigation.

Table 2: Subset of Mammalian genes with cCTEs captured in the Vector Trap and Sequenced on Oxford Nanopore Technology and MiSeq Platforms

Gene Name	ONT (reads)	MiSeq (reads)	CTE Alignment Location	Insert Sizes (bp)	Annotated IR hg38
TCEAL8	10787	1864	UTR3	633	No
P4HB	2483	357	Multiple Exons Multiple Exons+ UTR3	534 699	Yes
ACTN4	3293	445	Multiple Exons Multiple Exons	401 762	Yes
FAM96A	1606	379	UTR3	694	No
CRKL	3207	1	UTR3	566	No
COL4A2	8484	1250	Intron Intron	156 199	Yes
SEMA3C	2648	749	UTR3	773	No
KRT7	1695	2064	Exon+UTR3	306	Yes
CUX1	1061	56	Intron	665	Yes
YBX3	98	55	Multiple Exons + UTR3	901	Yes
EEF2	615	1316	Multiple Exons	719	Yes

EIF3G	857	614	Multiple Exons	758	Yes
GNAS	197	84	Multiple Exons + UTR3 Multiple Exons + UTR3	857 686	Yes
SDC4	372	92	Multiple Exons + UTR3 Exon +UTR3	599 528	No
PSMD3	310	488	Multiple Exons + UTR3	586	Yes
PDCD11	438	1510	Multiple Exons + UTR3	654	Yes
PSMB1	87	237	Multiple Exons + UTR3	816	No
CDC42BPB	14	233	Multiple Exons + UTR3	861	Yes
ANKRD1	303	323	Multiple Exons	224	No
CDV3	67	334	UTR3	786	No
MACF1	28	643	Multiple Exons Multiple Exons	427 476	Yes
TRIP13	413	977	UTR3	182	Yes
RPL18A	15	1	UTR3	532	Yes
CXCL3	46	158	Multiple Exons + UTR3	392	No
HEATR3	11	1	Multiple Exons	706	Yes
ACTB	34	4	Multiple Exons Multiple Exons	732 321	Yes
CHST1	205	513	Intron	230	Yes
TMEM242	271	833	UTR3	109	No
HSP90B1	34	8	Multiple Exons	690	Yes
KPNB1	68	124	Multiple Exons	764	Yes
VIM	19	259	Multiple Exons Multiple Exons	888 522	Yes
AP3D1	30	101	Multiple Exons	705	Yes
ATP11A	40	65	Intron	393	Yes
CCDC47	33	104	Multiple Exons	608	Yes
USP39	73	17	UTR3	321	Yes
ACTN1	8	0	Multiple Exons	693	Yes
RPS5	66	92	Multiple Exons + UTR3	726	Yes
CDC42SE2	131	416	Intron	227	No
SYT3	24	2	Multiple Exons + UTR3	731	Yes
CCT6A	216	3952	Multiple Exons	108	Yes
MDH2	26	59	UTR5 + Multiple Exons UTR5 + Multiple Exons	804 313	Yes
PPP6R1	15	0	Multiple Exons	852	Yes
SAMD4A	48	222	UTR3	565	Yes
MUS81	2	6	Multiple Exons + Intron	381	Yes
PRRC2A	1	0	Multiple Exons + UTR3	690	Yes
RPLP0	1	86	Multiple Exons Multiple Exons	439 147	Yes
LONP1	5	3	Multiple Exons + UTR3 Multiple Exons + UTR3	913 795	Yes
SIRT7	1	5	Multiple Exons	683	Yes

SMARCB1	0	2	Multiple Exons + UTR3	800	Yes
PYCR2	0	1096	Intron + Multiple Exons	1187	Yes
MXRA7	20	1252	UTR3	254	No
TMTC3	2	2	UTR3	227	No
NONO	2	23	Exon + UTR3	196	Yes

Abbreviations- UTR3- 3' untranslated region of gene, CTE- Constitutive Transport Element, IR-Intron Retention

As discussed above, the long-read sequencing platform allowed us to capture full cCTE elements and determine consensus sequences for these. Additionally, we were able to determine where the cCTE element aligned with respect to the specific gene architecture (e.g 3'UTR, intron or exon). For this, the long reads indeed proved critical as the full sequence of the cCTE allowed us to map across multiple exon-intron junctions and appropriately determine the architecture of the insert which we were usually unable to do with the shorter MiSeq reads. Aligned .BAM files for a subset of key genes with potential cCTEs were imported into the bioinformatics platform Geneious (Biomatters, Auckland, New Zealand). All reads across a particular genomic region were grouped. As long read sequences from ONT have a relatively high sequencing error rate compared to short read technology, we used all reads across a particular genomic region to form a consensus insert sequence. Of note, if more than one insert size was captured on the ONT platform across the same genomic region, the shortest functional sequence was used to determine the consensus insert sequence. In addition to the bioinformatic capability, Geneious has a robust genomic visualization platform. This allowed us to visually display where the CTE aligned within the gene architecture (alignment location for a subset of key genes are noted in Table 2). The visualization of genomic alignments from a subset of key cCTEs are shown in Figure 15.

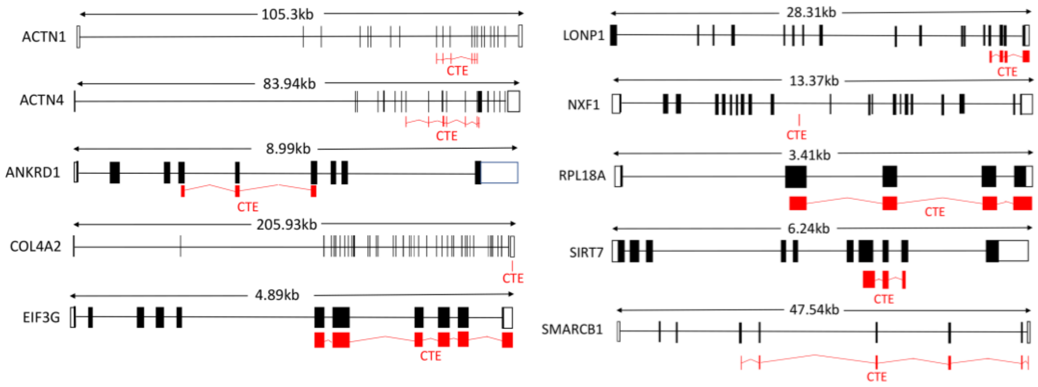

Figure 15: Cellular CTEs mapped to Gene Architecture. Following identification and sequencing of individual cellular fragments capable of mediating intron retention of an intron containing mRNA, we sub-selected cDNA fragments that were inserted in the sense direction with respect to the HIV vector (5' to 3' of respective mRNA). These fragments were then mapped to the human genome (Hg38) identifying which cellular gene contained the cellular fragment. Furthermore, we used the ONT long reads to determine the exact architecture of the cDNA fragment across the gene. The above diagram represents several key cellular CTE elements with the gene architecture of each gene in black including the length of the gene. The red bars highlight where the cDNA fragment or cellular CTE aligned across the gene.

Use of a Dual-Color Reporter Vector to Quantify the Functionality of Multiple cCTEs

To verify that a specific cCTE sequence was capable of mediating nuclear export and utilization of mRNA with a retained intron, we used a previously described dual-color HIV reporter assay (referred to in this thesis as Super Vector or SV), which has been previously described [173]. The HIV reporter has a green fluorescent protein gene sequence (eGFP) inserted into the gag region and a red fluorescent protein sequence (RFP- mCherry) inserted into the nef region (Figure 16, Panel A).

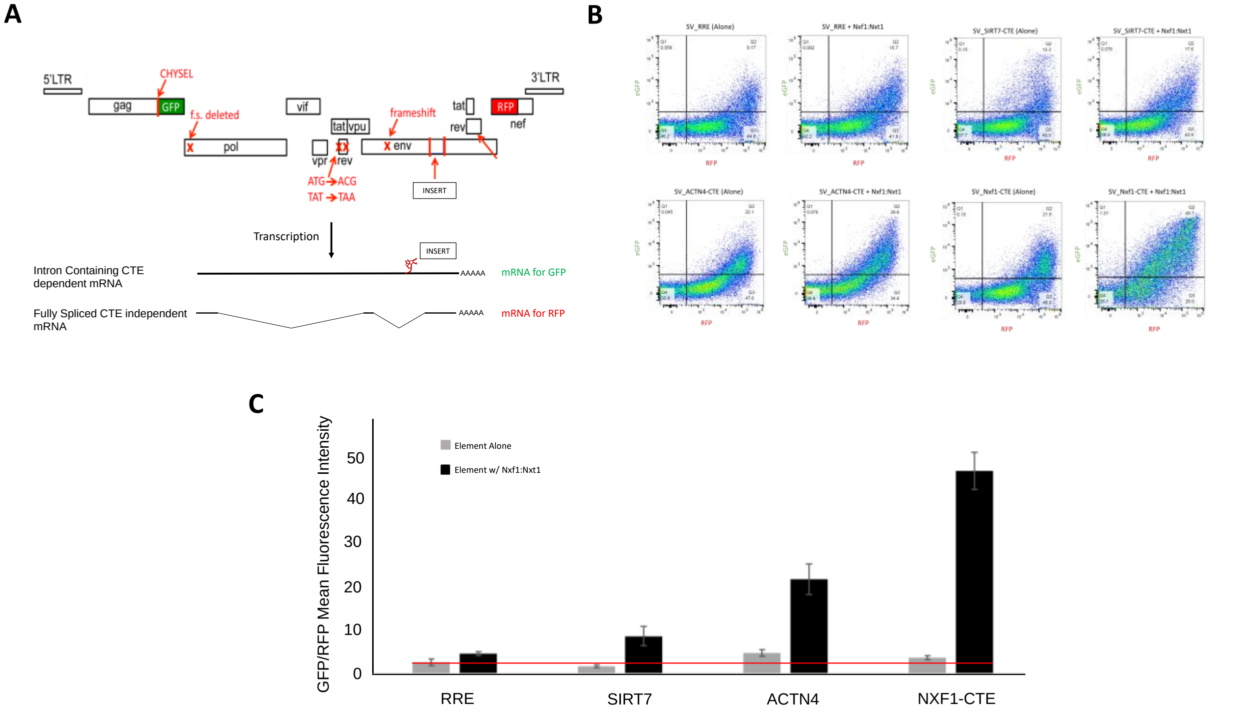

Figure 16: Dual Color Reporter Vector Measures Cellular CTE Function by Flow Cytometry. Panel A-The reporter assay used in this experiment is an HIV vector with a green fluorescent protein (GFP) sequence inserted into the gag region of HIV and a red fluorescent protein (RFP) sequenced inserted into the nef region of HIV. Nef is completely spliced and not dependent on a nuclear export signal, so RFP is constitutively expressed. Gag and thus GFP, remain dependent on a nuclear export signal. The supervector was designed to allow quick inserts of different nuclear export elements. In the experiment, 293T cells (4.5x10^5 cells in 6 well plate, 2 ml media) were transfected with 2,000 ng of the reporter containing various *cis*-acting elements (RRE, SIRT7-CTE, ACTN4-CTE and NXF1-CTE) together with and without 100 ng of pCMVNxf1, 200 ng of pCMVNxt1. Experimental conditions were performed in triplicate. Cells were harvested 48 hours post transfection and were immediately prepared for flow cytometry to generate a GFP to RFP mean fluorescent intensity ratio. Flow cytometry was performed on a Attune NxT flow cytometer with an attachment autosampler (Thermofisher Scientific). Data analysis was performed using FlowJo v10 (FlowJo, LLC). The gating strategies have been previously described (Jackson, 2020). Panel B- The mean fluorescent intensity (MFI) of eGFP (Q2) and mCherry (RFP) (Q3) for each individual cell was recorded. Functional activity of a particular CTE was determined by the ratio of eGFP to RFP. Panel C- The cis-acting elements from each report vector are located on the x-axis. The average value of eGFP/RFP MFI ratio is listed on the y-axis with standard deviation error bars. The value for each cCTE with and without Nxf1:Nxt1 cotransfection is listed. The red horizontal line is the ratio of RRE alone and represents the background level of the experiment.

The vector also has 2 mutations in the Rev gene that abolishes Rev function. In this vector, RFP is produced from a fully spliced mRNA that is not dependent on a special export element for expression. Thus, RFP is expressed constitutively from the reporter and is used to measure transfection efficiency. However, since eGFP is encoded within the gag intron, it is only produced if the appropriately paired nuclear export protein and cis-acting RNA element are present.

In a previous publication we demonstrated that two of the first discovered cellular CTEs in mammalian genes, ACTN4 and SIRT7, function in export and translation of a retained intron [27]. Additionally, we have unpublished Northern blot data in the Hammarskjold/Rekosh lab that show that both the ACTN4 and SIRT7 CTEs appear to be directly involved in nuclear export competency, as opposed to merely having effects on RNA stabilization. With both the ACTN4-CTE and the SIRT7-CTE, the levels of IR-RNA were unchanged in total RNA, while they were significantly increased in cytoplasmic RNA. Additionally, we have unpublished P24-ELISA data that show that ACTN4-CTE and SIRT7-CTE increases the amount of protein from an mRNA with a retained intron.

Given the above experimental data, I decided to demonstrate the appropriateness of the SV using the ACTN4-CTE and the SIRT7-CTE inserts in addition to the SV-RRE insert (without Rev as a negative control) and with an Nxf1-CTE insert as a positive control (demonstrated to have functionality in original methods publication) [173]. Each CTE containing plasmid and the control RRE plasmid was transfected with and without Nxf1:Nxt1, in triplicate independent transfections, to determine potential statistical significance. After 48 hours we harvested cells and performed flow-cytometry to quantify the mean fluorescent intensity (MFI) of both RFP and eGFP. The gating strategy and analysis for the flow cytometry data was completed as suggested in the methods publication of the Vector [173]. Additionally, as suggested by the publication, to control for transfection efficiency, we report the ratio of eGFP/RFP MFI across conditions (example of flow data for each condition is provided in Figure 16, Panel B, note each experimental condition was performed in triplicate though only one transfection flow data is displayed).

In both transfection conditions (alone, and with the addition of Nxf1:Nxt1), the RRE demonstrated a very low GFP/RFP ratio (Figure 16, Panel C). These results are consistent with previously published investigations as the RRE is dependent on Rev for any meaningful nuclear export, and Rev was not provided in the experiment. The RRE was used as both the negative control for the experiment and the baseline level of function for comparison with other export elements. We repeated the experiment with the ACTN4-CTE insert, SIRT7-CTE insert and both Nxf1-CTE inserts (same transfection conditions). ACTN4-CTE demonstrated a 1.8-fold change above the RRE at baseline ($p = 0.0003$). ACTN4-CTE also increased the eGFP/RFP ratio with the addition of Nxf1:Nxt1 to 4.6-fold above the level of the element alone ($p < 0.0001$). When SIRT7-CTE was transfected without additional proteins it did not raise the GFP/RFP ratio above the RRE background level. However, with the addition of Nxf1:Nxt1, the eGFP/RFP ratio increased 5.3-fold from that of the element alone ($p = 0.0005$). The Nxf1-CTE demonstrated a 1.4-fold change above the RRE baseline. However, when Nxf1:Nxt1 was added to the transfection experiment, the Nxf1-CTE demonstrated a 12.9-fold increase above the element at baseline ($p < 0.0001$).

This initial experiment in the dual color reporter assay utilizing the ACTN4-CTE, SIRT7-CTE, NXF1-CTE inserts were largely consistent with the published and unpublished northern blot (RNA nuclear export efficiency) and P24 (translation efficiency) experiments. The ACTN4-CTE and the SIRT7-CTE elements alone (without addition of Nxf1:Nxt1) demonstrated functionally at or just above the RRE baseline. However, with the addition of the nuclear export proteins (Nxf1:Nxt1) the MFI of eGFP/RFP ratio increased significantly with both elements. The data from the dual reporter assay clearly redemonstrates that a paired cis-acting nuclear export signal and the corresponding nuclear export protein are required for effective nuclear export of intron containing mRNA in eukaryotic cells.

Multiple cCTEs effectively mediate expression of mRNA with retained introns

We next selected a subset of approximately 20 genes with cCTEs identified in the vector trap (Table 3). The elements were selected based on 2 criteria. The first criteria were elements that demonstrated increased frequency in the Vector Trap. The second criteria were elements in genes with significant evidence of intron retention, based on an extensive literature review. Each cCTE sequence was inserted in-place of the RRE in the HIV dual-color reporter (Super Vector). The resulting plasmids were transfected into 293T cells with and without Nxf1:Nxt1 co-expression. The cells were harvested at 48 hours (all experiments performed in biologic triplicate) and analyzed by flow cytometry. For all cCTEs, eGFP/RFP ratios were statistically compared to either the baseline RRE eGFP/RFP ratio or were compared to the eGFP/RFP ratio of the element with and without the addition of Nxf1:Nxt1. Statistical significance was determined by a two-sided Student T-test (using the geometric mean of the eGFP/RFP ratio in triplicate). A 1.5-fold increase above baseline measure with a p-value < 0.05 was considered significant. In these experiments, the reporter vector containing the RRE without Rev was used as a control to determine the background levels of eGFP.

Table 3: Mammalian genes with cCTEs identified in Vector Trap and sequenced on ONT platform tested alone and with the addition of Nxf1 and Nxt1

CTE	**GPF: RFP MFI Element (SD)**	**Fold Change (RRE)**	**P-value**	**GFP: RFP MFI +NN (SD)**	**Fold Change (+/- NN)**	**P-value**
Experimental Controls						
HIV_RRE	2.6 (0.7)	1	--	4.5 (0.4)	1.7	P < 0.0001
MPMV-CTE	8.0 (2.6)	3.1	P < 0.0001	58.7 (13.2)	7.3	P < 0.0001
NXF1-CTE	3.6 (0.4)	1.4	P < 0.0001	46.6 (4.3)	12.9	P < 0.0001
Cellular CTEs tested in Super Vector						

EIF3G	37.0 (10.6)	14.4	P = 0.0001	49.9 (15.9)	1.3	P = 0.1035
ACTN4	4.7 (0.8)	1.8	P = 0.0003	21.7 (3.5)	4.6	P < 0.0001
ACTN1	14.4 (5.0)	5.6	P = 0.0057	34.1 (5.1)	2.4	P = 0.0003
HSP90B1	14.7 (6.4)	5.7	P = 0.0323	24.5 (4.3)	1.7	P = 0.0467
LONP1	15.9 (3.2)	6.2	P = 0.0034	44.5 (20.6)	2.8	P = 0.0347
MUS81	3.6 (0.7)	1.4	P = 0.0517	5.7 (0.2)	1.6	P = 0.0064
ANKRD1	2.2 (0.1)	0.8	P = 0.072	6.0 (0.3)	2.7	P = 0.0009
RLP18A	4.4 (0.3)	1.7	P = 0.0003	17.3 (5.5)	3.9	P = 0.0549
RPLP0	3.0 (0.04)	1.2	P = 0.0360	4.9 (0.2)	1.6	P = 0.0012
TSR2	3.2 (0.2)	1.2	P = 0.0105	4.8 (0.4)	1.5	P = 0.0124
XRN2	6.8 (0.3)	2.7	P < 0.0001	12.6 (0.4)	1.8	P < 0.0001
COL4A2	5.3 (1.1)	2.0	P = 0.0168	5.0 (0.2)	1.0	P = 0.8378
CRKL	5.3 (0.2)	2.1	P < 0.0001	8.2 (0.6)	1.5	P = 0.0627
PRRC2	4.5 (0.2)	1.7	P < 0.0001	5.1 (0.1)	1.1	P = 0.0893
TCEAL8	4.3 (0.2)	1.7	P = 0.0002	5.1 (0.3)	1.2	P = 0.1347
SIRT7	1.7 (0.3)	0.6	P = 0.0012	8.6 (2.2)	5.2	P = 0.0005
SMARCB1	9.7 (0.2)	3.8	P < 0.0001	20.0 (4.4)	2.1	P = 0.0548
PYCR2	3.9 (0.2)	1.5	P = 0.0001	42.0 (3.0)	10.8	P = 0.0019

All experimental conditions were performed in triplicate. Standard deviation is listed in parathesis for all GFP/RFP ratios. Statistical significance was determined by two-sided Student T-test. The 4th column represents a statistical comparison of the element alone compared to the RRE alone (which represents the experimental background of the vector. The last column represents a statistical comparison between the element alone and the element with Nxf1:Nxt1 co-transfection. Abbreviations- RRE- Rev Response Element, NN- Nxf1:Nxt1, GFP- Green Fluorescent Protein, RFP- Red Fluorescent Protein, SD- Standard Deviation.

A subset of cCTEs (13 of 23) demonstrated significant eGFP expression (as compared to the RRE) when plasmids with the elements were transfected alone (Figure 17, Panel A). However, the efficiency of the elements varied considerably. Some elements, such as PYCR2-CTE and TCEAL8-CTE demonstrated low function with levels only 1.5-fold over those from the RRE-containing plasmids. Other cCTEs, such as XRN2-CTE and SMARCB1-CTE showed levels that were similar to the MPMV-CTE. Additionally, 4 elements- ACTN1-CTE, HSP90B1-CTE and LONP1-CTE demonstrated at least a 5-fold increase in GFP/RFP MFI ratio above the RRE alone. EIF3G-CTE gave a 14-fold increase over the RRE and was the most efficient element as determined by this assay.

Many of the elements (12 of 23) demonstrated increased functionality with the addition of Nxf1:Nxt1 (Figure 17, Panel B). Some elements, including RPLP0-CTE and MUS81-CTE showed a fold-increase between 1.5 and 2. Other elements, such as PYCR2-CTE, had a greater than 10-fold increase above the element baseline when Nxf1:Nxt1 was transfected. In fact, the PYCR2-CTE, baseline function with endogenous Nxf1 (Nxf1 level expressed in 293Tcells) was similar to the of Nxf1-CTE and the response to over-expression of Nxf1:Nxt1 was also similar to what was seen with the Nxf1-CTE.

Some elements, such as the ACTN1-CTE demonstrated both moderate baseline functionality (3.5-fold increase over RRE) as well as increased function with co-transfected Nxf1:Nxt1 (2.3-fold).   There were other elements, such as that from EIF3G that demonstrated a very strong baseline functionality, but did not respond to the addition of Nxf1:Nxt1, suggesting the element may not be dependent on Nxf1 for function. Of the 23 elements tested, 5 elements did not show significant functionality in the dual-color reporter assay (Figure 17, Panel C).   For example, the CCT6A-CTE had nearly 4,000 reads on the MiSeq sequence run. However, the element's baseline function was equivalent to background RRE and did not show significant response to Nxf1:Nxt1. It is possible that this was a result of not isolating the correct sequence (error in trimming the 3' or 5' end of the CTE sequence) or that the secondary structure was altered in the mRNA expressed from the HIV reporter.   It should also be noted that the vector trap specifically selected for elements that were capable of nuclear export of intron containing mRNA. However, selection did not require that the CTEs were also capable of promoting utilization in the cytoplasm. CTEs that only promoted export would not score in the HIV reporter assay, since this requires both export and translation of the IR mRNA.

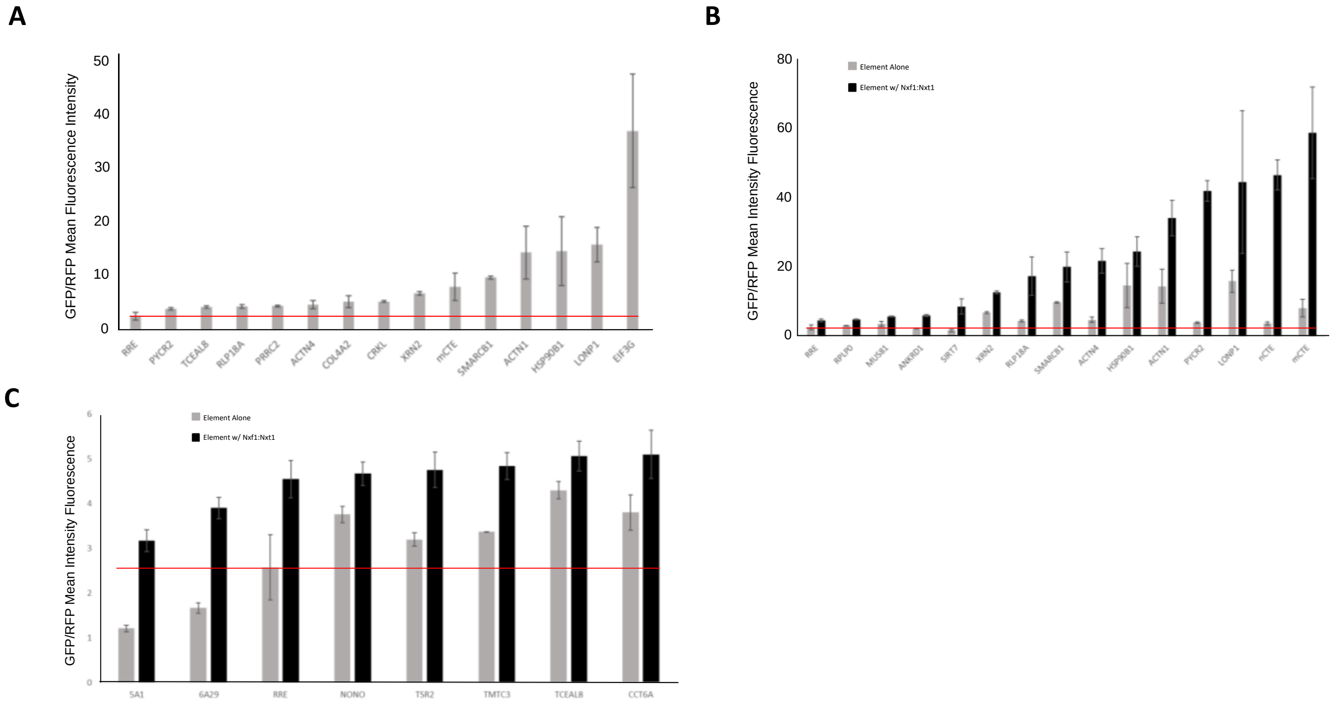

Figure 17: Cellular CTEs tested in Supervector demonstrated variable functionality above experimental background with and without the addition of Nxf1:Nxt1. 293T cells (4.5x105 cells in 6 well plate, 2 ml media) were transfected with 1,000 ng of Supervector with the respective cellular CTE inserted. All experimental conditions were completed in triplicate. At 48 hours, cells were harvested and underwent flow cytometry to calculate the eGFP and RFP mean fluorescent intensity. The individual CTEs found in each respective gene are located on the x-axis. GFP/RFP MFI ratio is listed on the y-axis. The value for each cCTE represents the average of 3 independent transfections. Error bars represent the standard deviation. The red line is the eGPF/RFP ratio of the RRE element alone and represents the background expression level of the assay and is displayed in 3 figures. A series of elements demonstrated statistically significant increase in expression above RRE background (Panel A). Another set of elements demonstrated significant increase with the addition of exogenous Nxf1:Nxt1 (Panel B). There were also a set of elements that did not show functionality in the assay (Panel C).

Many cCTEs are found in genes with known intron retention and cCTE containing genes cluster in cellular processes related to post-transcriptional regulation and neural differentiation

From the total cCTE list, there was a clear over-representation of elements in genes in the ribosomal binding protein family (RPL and RPS), cytoskeletal genes, and zinc finger binding protein (ZNF) genes. Given the possible convergence on specific cellular pathways, we performed a formal Gene Ontology Gene Enrichment Analysis (Biological Process, Cellular Component and Molecular Function) on genes with cCTEs discovered in the vector trap. From the full list of potential cellular CTEs which mapped to genes, we ran a GO enrichment analysis using the clusterProfiler package in R and ShinyGO platform for visualization [151, 174]. Gene pathways with an enrichment false discovery rate < 0.05 were included in the analysis. Interestingly, 10 of the top 20 functional categories in the biological process pathway were directly involved in neuronal development and differentiation (Figure 18). Other key functional categories represented in the top 20 functional pathways included cellular localization, cell morphogenesis, regulation of development processes and regulation of cell differentiation. Additionally, 9 of the top 10 cellular pathways in the molecular function category were directly involved in RNA or general nucleotide binding.

cCTEs were discovered in many genes, which have previously been shown to have IR protein isoforms. For example, a functional cCTE was discovered in the ANKRD1 gene. ANKRD1 (also known as CARP) produces multiple alternatively spliced transcripts, 3 of which contain introns (variable inclusion of introns 6, 7 & 8) [175]. These IR isoforms have been shown to be efficiently exported to the cytoplasm and translated into proteins in myocardium [176]. Two other cCTEs were discovered in the genes HOOK2 and YBX3, which have also been shown to have stably maintained IR isoforms in the nucleus during spermatogenesis and undergo delayed export after approximately 9 days [177]. This could represent a separate class of cCTEs, where the CTE may serve to stabilize the RNA in the nucleus, followed by nuclear export when a specific export factor becomes available.

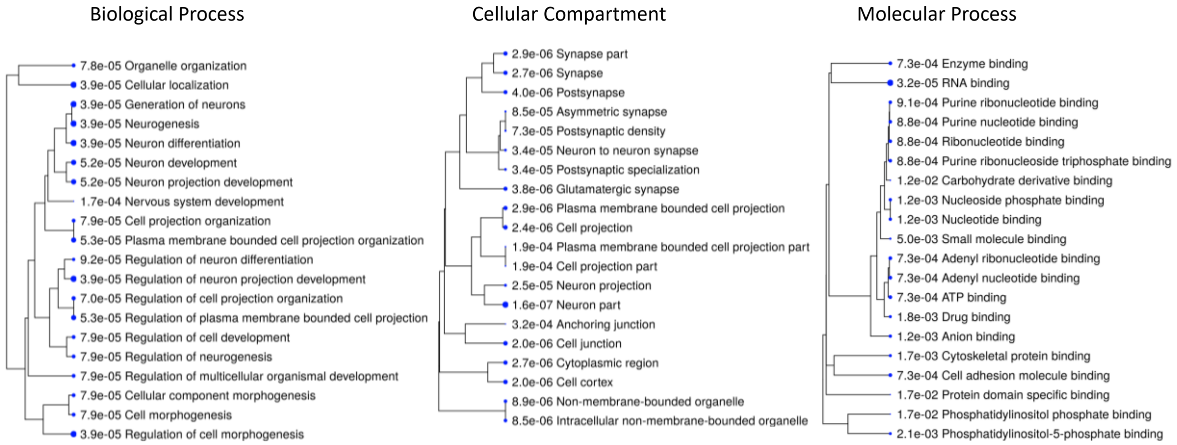

Figure 18- Gene Ontology Enrichment Analysis of cellular CTEs demonstrate a role of CTEs in neuronal differentiation and RNA binding. Potential Cellular CTEs that were identified in the vector trap and sequenced on either the MiSeq or ONT platform were identified. A Gene Ontology Enrichment Analysis was performed using the ShinyGO visualization platform. Gene pathways that demonstrated an enrichment false discovery rate < 0.05 were included. Displayed above are the top 20 gene pathways (organized by FDR significance) in each respective category of the GO analysis- Biological Process, Cellular Component and Molecular Process.

Over-Expression of Nxf1/Nxt1 increases Intron Retention in Cytoplasmic mRNA from Multiple Genes

Nxf1:Nxt1 Overexpression RNA-seq experiment and Intron Retention Analysis Methodology

All mRNAs require export from the nucleus to the cytoplasm through the nuclear pore. This is a tightly regulated process that appears to be directly coupled to transcription, alternative splicing and nuclear RNA processing [178]. The canonical export pathway of bulk mRNA is generally understood to be a co-transcriptional binding of the TREX (Transcript Export) Complex and the recruitment of the Nxf1:Nxt1 dimer which facilities transition across the nuclear pore [179, 180]. However, this data has been largely influenced by early studies of Drosophila S2 cells [181] and yeast [182]. Multiple investigations in mammalian cells have demonstrated weak direct affinity of Nxf1 for the majority of mRNA [183, 184], leading to numerous studies into potential adapter proteins that may promote Nxf1 RNA binding, including ALY (a TREX component) [185] and SR proteins [186-188]. Though it is still generally assumed that both Nxf1 and TREX have complimentary roles in mRNA export, recent iRNA knockdown investigations have demonstrated non-uniform effect on mRNA subsets with siRNA against Nxf1 and TREX [189]. These results highlight that the nuclear mRNA export pathways remain incompletely understood and indicate that Nxf1 may demonstrate mRNA export selectivity. Additionally, our data show that Nxf1 can directly bind RNA with specific RNA secondary structures (CTEs) to facilitate nuclear export. Data from our lab also suggest that Nxf1 plays a key role in the cytoplasmic utilization of mRNA and more efficient association with ribosomes [161]. This is also consistent with recent super-resolution microscopy data which suggests a key role of Nxf1 on the cytoplasmic side of the nuclear pore [190].

Given the data shown above that increased expression of Nxf1:Nxt1 significantly promotes the function of many CTEs, as well as increasing data that Nxf1 may demonstrate selectively in RNA export, we choose to pursue a global RNA-seq experiment to discover if over-expression of Nxf1:Nxt1 increased cytoplasmic intron retention events. We transfected vectors expressing Nxf1 and Nxt1 (1:2 ratio) into 293T cells and at 48 hours post-transfection we isolated both total and cytoplasmic RNA. For an experimental control, we also transfected a pCMV-empty vector (EV) plasmid. All experiments were performed in triplicate for statistical analysis. Isolated RNA (both total and cytoplasmic) was polyA selected and sent to Novogene for short read RNA-sequencing (stranded 150 bp pair-end reads, ~40 million read depth).

Interestingly, when we performed a standard differential gene expression pipeline using the HiSat2, Stringtie and Ballgown pipeline between cytoplasmic Nxf1:Nxt1 overexpression compared to cytoplasmic Empty Vector Control, there were only 46 differentially expressed genes. The results suggests that Nxf1 is not affecting transcription or bulk nuclear export quantified at the gene level. However, as discussed above, differential gene expression pipelines do not capture differential intron retention, other alternative splicing events or other events that result in mRNA isoform changes. Because of our specific interest in IR, I first decided to analyze the RNA-seq data for differential intron retention events.

Analysis of the RNA-seq data for intron retention discovery continues to be a computational challenge. There currently is not a best practice standard bioinformatic pipeline for intron retention discovery using RNA-seq data. This is in part due to the limited number of dedicated bioinformatic tools to evaluate IR, as well as the intrinsic difficulty in accurate assessment of intron expression [38]. Key aspects of IR analysis include whether to account for overlapping features including differentially spliced transcripts, how to overcome sequencing and alignment artifacts including repeat and low complexity regions and correcting for low coverage of flanking exons and 3' coverage bias [9, 36, 37, 45]. IR levels are often estimated using one of two measures: the percentage spliced in (PSI) or the IRratio. The PSI is frequently used to measure alternative exon splicing [191] and recently an intronic PSI has been adopted as the level of transcripts (in transcripts per million (TPM)) supporting the retention of the intron, compared to the level of transcripts supporting its exclusion [36]. IRratio is the ratio of intronic reads compared to the immediately spliced exon reads [9]. Both of these ratios tend to show high fluctuations and their behavior is difficult to model statistically. As a consequence, no approach has been developed to estimate dispersions and confidence intervals. Because IR detection requires the aforementioned corrections and specific measurements, we utilized several bioinformatic approaches to measure IR in our Nxf1 RNA-seq experiment. Of note, in collaboration with a group at the University of Montpellier, our group recently published a best practice review of intron retention discovery using RNA-seq data, as well as a new tool for IR detection using 3rd generation long reads sequencing [38].

For robustness of IR discovery in our RNA Seq data, I used 4 separate bioinformatic pipelines with different sensitivities and computational assumptions. I used a standard splice aware differential transcript usage pipeline- HISAT2 (read alignment), Stringtie (transcript assembly), Ballgown (transcript/intron quantification in R) and differentially quantified all reads that aligned within introns between conditions (pipeline did not take into consideration either junction reads or 5' and 3' exon usage) [154]; I utilized a pseudoalignment (Salmon) pipeline and quantified differential transcript expression of transcripts with annotated retained introns in Ensembl [147]; I utilized rMATS (robust Multivariate Analysis of Transcript Splicing) which is a differential splicing quantification tool that measures alternative splicing using a retained intron PSI ratio [192]; and IRFinder- a dedicated intron retention program based on STAR alignment and IRratio quantification between samples [9]. Significant IR events were considered if absolute value of the log2fold change was > 1.5 and the false discovery rate (p-adjusted value) was < 0.05. I then did multiple list comparisons between the IR events calculated from each program and used IGViewer to visually inspect top intron retention candidates.

Nxf1:Nxt1 overexpression increases intron retention in cytoplasmic mRNA and affects cellular processes involved in RNA binding and RNA splicing

Each bioinformatic approach led to the discovery of numerous statistically significant intron retention events when comparing Nxf1:Nxt1 overexpression to EV control in isolated cytoplasmic RNA. As suspected the Hisat2-Stringtie-Ballgown pipeline, which quantified the normalized read counts across introns and then calculated a differential intron usage between conditions (similar to differential exon usage) was the least specific. There were 2,272 introns that were differentially expressed between conditions, and 1,816 of these were significantly increased in Nxf1:Nxt1 over-expression (80%) (Appendix Table 2). When I performed a differential transcript expression analysis between Nxf1:Nxt1 using Salmon and DeSeq2 we noted 1,691 significant RNA isoform changes. By comparing this list to a list of annotated retained introns in Ensembl, we noted that 434 of these were annotated as retained introns (26%) (Appendix Table 3). Of the 434 retained intron RNAs, 280 were increased by Nxf1 expression (65%).

I used rMATS to calculate the percentage spliced in (PSI) of intron containing splicing events between Nxf1:Nxt1 and EV, which found 1,409 significant intron retention events (Appendix Table 4). When I constrained the results to the recommend PSI levels between absolute value of 0.05 and 0.7, there were 672 independent genes with significant intron retention events of which 361 demonstrated upregulated intron retention (54%). Lastly, I performed a differential intron retention pipeline between cytoplasmic Nxf1:Nxt1 over-expression and cytoplasmic EV using IRFinder. In this case there were 385 independent genes which demonstrated significant intron retention between conditions, 265 of which were upregulated in Nxf1:Nxt1 (70%) (Appendix Table 5). Of note, ballgown, like the salmon pipeline, remains limited by the annotation of transcript isoforms. For example, from the 2,272 mRNA introns that were differentially expressed in the ballgown pipeline- 523 IR mRNA (approximately 25%) were not associated with an annotated transcript making evaluation of biological significance challenging. This is an additional limitation to relying on previous transcriptional annotations for discovery of novel transcriptional events such as intron retention or proviral expression.

We performed a formal comparison of the significant intron retention events determined by the different intron retention analyses pipelines (Figure 19). At the gene level, there was significant though mild convergence of specific genes. Between the Ballgown (2,200) and rMATS (672) intron retention lists there were 28 genes that overlapped (Fisher exact test, p < 0.0001), between Ballgown and IRFinder there were 91 genes that overlapped (Fisher exact test, p < 0.0001). The convergence between lists was more significant with the more specific intron retention tools. The overlap of significant intron retention events between rMATS (672) and IRFinder (385) was 116 genes (Fisher exact test, p < 0.0001) which represents a 30% overlap. There were 28 genes that overlapped between the Ballgown, rMATS and IRFinder intron retention lists and there were 9 genes that overlapped all 4 intron retention lists.

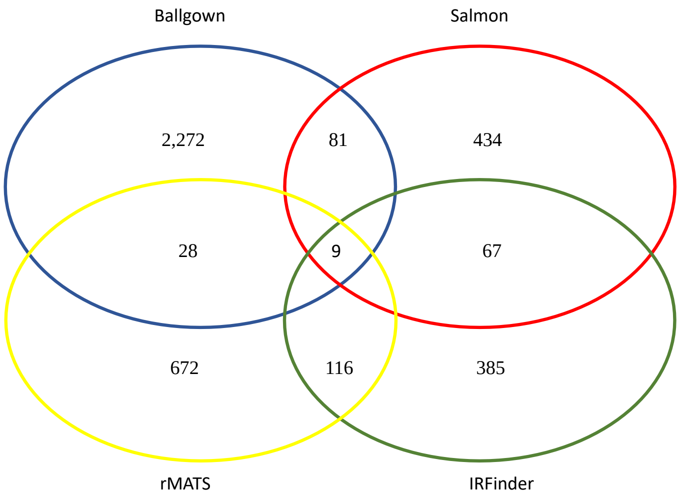

Figure 19- Overlap of Genes with Intron Retention Discovered in the Separate Intron Retention Computational Pipelines. Genes that demonstrated significant intron retention lists in each computational pipeline following comparison of cytoplasmic Nxf1:Nxt1 mRNA to cytoplasmic Empty Vector Control. The blue circle represents the Ballgown pipeline of differentially expressed intron reads; the red circle represents the Salmon pipeline of differential expressed transcripts annotated as retained introns in Ensembl; the yellow circle the rMATS pipeline for significant differentially retained introns based on a percentage spliced in calculation; and the green circle represents the IRFinder pipeline of retained introns calculated based on a IRatio calculation.

Similar to our original CTE findings, the significant intron retention lists generated from each bioinformatic pipeline had an apparent predominance of ribosomal proteins, splicing factors and cell cycle specific genes. For example, the top 3 annotated genes form the Ballgown analysis of differential intron usage demonstrated increased intron retention in mRNA from several ribosomal protein genes (RPL37, RPL15, RPS17) with significant fold changes of 287, 259 and 141, respectively. Additionally, ACTN4 demonstrated increased intron retention (1.7-fold), and ALYREF demonstrated increased intron retention (2.1-fold) in the Salmon pipeline. In addition to the noted convergence of specific genes with intron retention between bioinformatic pipelines, there was a stark convergence on biological pathways associated with Nxf1 mediated intron retention between all pipelines. We utilized Gene Ontology Enrichment Analysis (Biological Process, Cellular Component, Molecular Function) to analyze all intron retention lists generated from cytoplasmic mRNA analysis of Nxf1/Nxt1 overexpression (Figure 20, Panel A-D). The top gene enrichment lists form each bioinformatic pipelines converged on biological process associated with RNA splicing, RNA processing and mRNA metabolic processes. The cellular components pathways converged on the nuclear lumen, the ribosome and the spliceosome, whereas the Molecular Process Enrichment Analysis demonstrated RNA binding and mRNA binding in all lists.

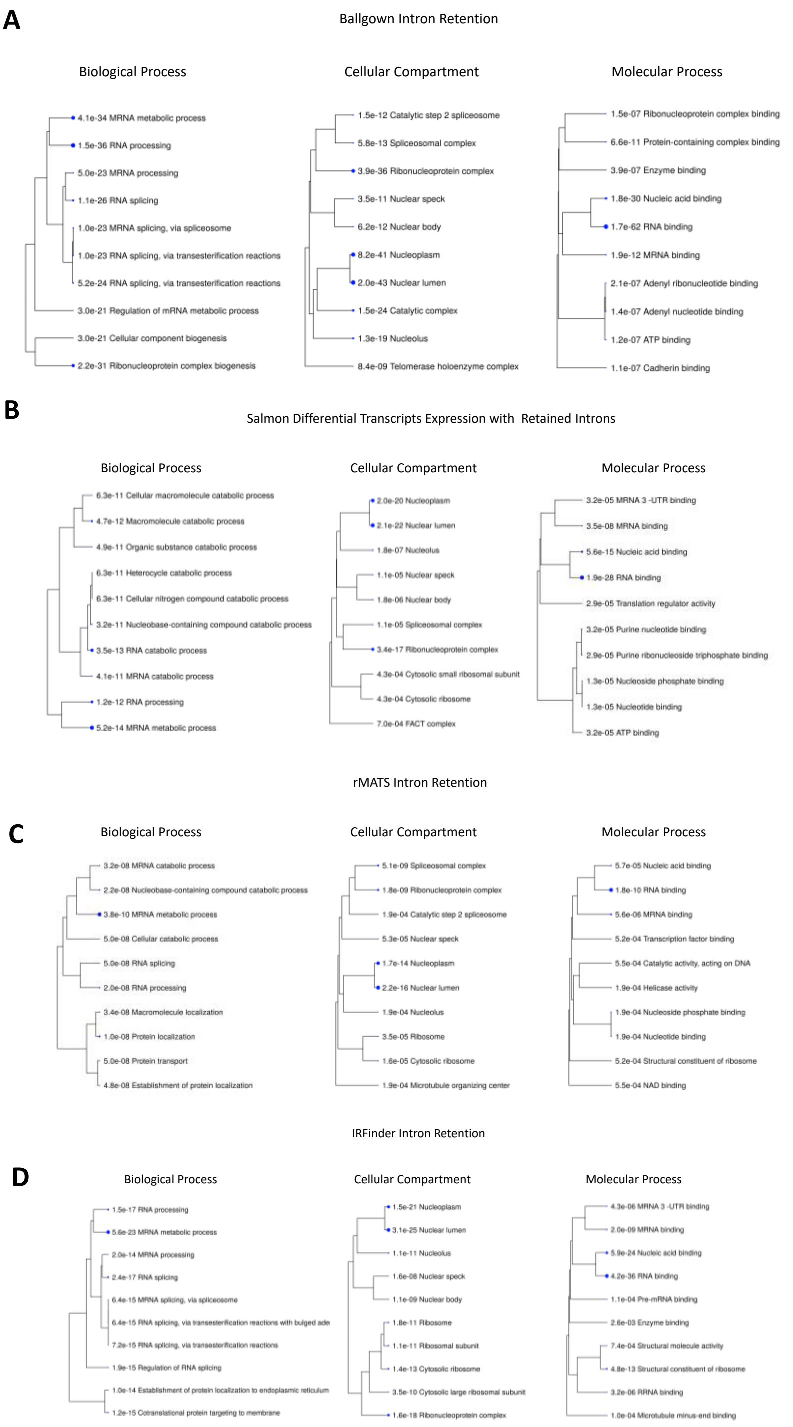

Figure 20: Gene Ontology Enrichment Analysis of Nxf1 Mediated Intron Retention Demonstrates Role in RNA processing, Spliceosome Complex and ribonucleoprotein complex binding. Genes with significant intron retention discovered in the (A) Ballgown pipeline (B) Salmon Differential Transcripts with Retained Introns pipeline (C) rMATS pipeline or (D) IRFinder pipeline following comparison of cytoplasmic Nxf1:Nxt1 overexpression with cytoplasmic EV control were identified. _A Gene Ontology Enrichment Analysis was performed using the ShinyGO visualization platform. Gene pathways that demonstrated an enrichment false discovery rate < 0.05 were included. Displayed above are the top 10 gene pathways (organized by FDR significance) in each respective category of the GO analysis- Biological Process, Cellular Component and Molecular Process.

in total EV to cytoplasmic EV which noted 14, 676 differentially expressed introns, 14,605 of

Nxf1:Nxt1 overexpression does not result in significantly increased intron retention in total mRNA

As describe above, we also isolated and performed Illumina short read sequencing on total mRNA from cells transfected with Nxf1:Nxt1 and an EV control. Total RNA was isolated from replicate plates at the same time as the cytoplasmic RNA. Given the IRFinder was the most selective of the intron retention pipelines, I utilized this pipeline for the subsequent analysis of the total RNA data. I first used IRFinder to compare intron retention between cellular compartments of Nxf1:Nxt1 overexpression (between total and cytoplasmic mRNA, polyA+). There were 9,013 differentially expressed introns between total and cytoplasmic compartments with Nxf1 overexpression, 8,971 of which were increased in total RNA (99.5%). This finding was duplicated when we compared intron retention between total EV data and cytoplasmic EV data which noted 14,676 differentially expressed introns, 14,605 of which were increased in total RNA (99.5%). This data suggests that indeed there is a large pool of mRNA with retained introns in nuclear mRNA compared to cytoplasmic mRNA, presumably due to nuclear export restriction at the nuclear pore. I then compared the intron expression pattern of total Nxf1:Nxt1 to total EV and found that there were only 189 significant intron retention events between conditions and only 31 introns that were increased with Nxf1:Nxt1 expression (16%) (Appendix Table 6)). Though the data is not conclusive, it clearly supports that Nxf1 selectively exports a specific subset of mRNA with retained introns in mammalian systems.

Overlap between genes that show an increased presence of cytoplasmic IR isoforms with Nxf1:Nxt1 overexpression and genes with cellular CTEs

Data from our CTE analysis suggests that Nxf1/Nxt1 promotes the function of numerous cellular CTEs. Thus, I decided to analyze if there was any correlation between genes with cellular CTEs and genes that demonstrated differential intron retention as a result of Nxf1/Nxt1 over-expression. I again used several of the intron retention lists generated from the different bioinformatic pipelines, specifically the most rigorous intron detection pipelines (rMATS and IRFinder). In comparison between the CTE list and the IR results from rMATs there were 37 overlapping genes (Fisher Exact test, p < 0.0001) (Figure 21, Panel A) while in comparison to the IRFinder results there were 30 overlapping genes (Fisher's Exact test, p < 0.0001) (Figure 21, Panel B). Several key genes that came up in multiple intron retention lists that also had cCTEs were HOOK2, LONP1, EIF3G, MUS81, LYPLA2, RPL18A, and ACTN4. As presented in the CTE discovery section of this chapter, the majority of these elements were tested in the supervector and were functional with the addition of Nxf1:Nxt1 (the only element not tested was HOOK2). The only exception of a tested element that did not show a significant increase with the addition of Nxf1/Nxt1 SV experiment was EIF3G. However, this element had a very high baseline functionality in 293T cells by itself.

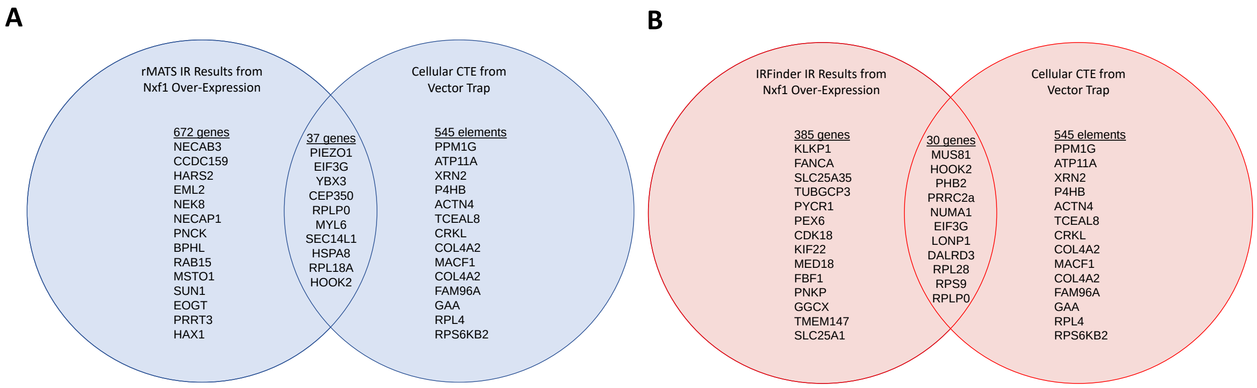

Figure 21, Panel A- Overlap between genes with cellular CTEs discovered in the vector trap and genes with differentially expressed cytoplasmic intron retention events calculated from the (Panel A) rMATS or Panel (B) IRFinder computational pipelines following Nxf1:Nxt1 overexpression compared to Empty Vector Control. Figures represent a subset of all overlap.

Long Read Sequencing following Nxf1:Nxt1 overexpression confirms an increase in mRNA with complete introns in the cytoplasm

Given the previously discussed limitations of bioinformatics analysis of IR, one of the key principles of intron retention analysis remains visual confirmation of IR events for key transcripts of interest [38]. Another biological principle of intron retention discovery using RNA-seq based analysis is confirmation that the retained intron is indeed a true molecular transcript of the gene and not a result of errant mapping of repeat or low complexity regions. This can be assured by demonstrating mRNA reads which contain the upstream 5' exon, the whole intron including 5' and 3' splice sites and the downstream 3' exon. This is possible with 3rd generation long read sequencing. We thus decided to use Oxford Nanopore Technology to sequence total and cytoplasmic RNA isolated from the Nxf1:Nxt1 and EV transfections. The sequencing libraries were prepared using a "direct cDNA" library kit, to minimize errors (#SQK-DCS109).

Following sequencing, we received approximately 5M reads per flow cell (2 flow cells utilized- one for cytoplasmic RNA, one for total RNA), which amounted to roughly 2.25M reads per experimental condition. The average PHRED quality score was 10.1 and the average read length was 1,337 bp, which is roughly consistent with the average length of protein coding mRNAs in humans. We then utilized the Pinfish bioinformatic pipeline as developed by ONT to process the long-read data [122]. Key steps to this bioinformatic analysis included selecting full-length reads using the program Pychopper, alignment of long reads to the human genome using Minmap2 (ONT mode) and visualization of aligned reads in IGViewer.

We utilized the IR differential analysis from the short read data to sub-select a small group of potential intron retention events to manually analyze using the long read ONT cytoplasmic mRNA data. We concentrated on genes that had direct overlap between the CTE list as well as either the rMATs or the IRFinder intron retention lists. Figure 22, Panel A is a graphical representation of the gene architecture of MUS81 that demonstrates where the CTE maps in the gene as well as where the IR event was observed with Nxf1/Nxt1 over-expression. Figure 22, Panel B represents the IGViewer output form the cytoplasmic Illumina short read data from all 3 Nxf1:Nxt1 over-expression samples and all 3 EV control samples across the retained intron (3' end of the MUS81 gene). As demonstrated by the read density- there is clearly increased read counts in the Nxf1:Nxt1 samples overlying the intron as compared to the EV control where there is essentially zero read density. Additionally, it is clear that the read density spans the entire intron as would be appropriate for a true intron retention event. Figure 22, Panel C is the IGViewer output of the same region of MUS81 with the long-read data from the cytoplasmic Nxf1:Nxt1 and cytoplasmic EV data. The long read data confirms the short read sequencing data and clearly demonstrates multiple examples of a complete mRNA read spanning from the 5' exon through the intron into the 3' exon (including both the 5' splice site and the 3' splice site) with Nxf1:Nxt1 overexpression.

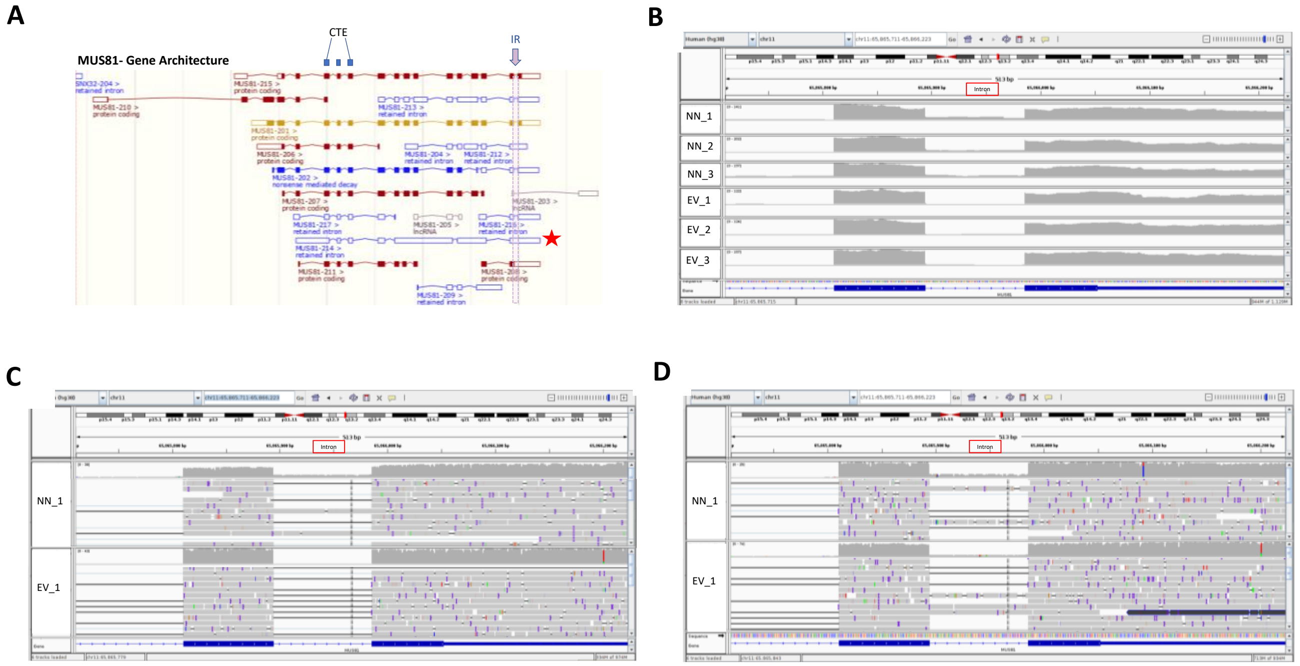

Figure 22: Nxf1:Nxt1 increases cytoplasmic intron retention in MUS81. Panel A represents a graphical description of the MUS81 gene architecture from Ensembl. The location of the cellular CTE that mapped to MUS81 is located in blue at the top of the figure. The retained intron discovered from the Nxf1:Nxt1 over-expression experiment is located in purple (3' end of the gene). The red star denotes the annotated transcript that contains both the CTE and the retained intron. For both the short-read and long read (panel C) experiment, 293T cells (10 cm plates, $3x10^6$, 10 ml media) were transfected with 5 ug of pCMVNxf1 and 10 ug or pCMVNxt1 or with 15 ug of pCMV-Empty-Vector. Cells were harvested 72 hours post-transfection. Each experimental condition was performed in triplicate. Cytoplasmic mRNA and total mRNA were isolated using a Phenol: Chloroform protocol. Short read data was sequenced on an Illumina HiSeq platform (stranded 150 bp paired-end reads). Following transfection and cytoplasmic mRNA isolation, equimolar amounts of RNA from each triplicate experiment was pooled for a total of 50 ug of polyA selected RNA for each condition (Nxf1:Nxt1 and Empty Vector). A direct cDNA native barcoding ONT protocol was used to sequence the RNA on the ONT platform. Panel B represents the mapped cytoplasmic mRNA short reads across the 3' end of the MUS81 gene. Panel C represents the mapped cytoplasmic mRNA long read data across the same gene region. Panel D represents the mapped total mRNA long read data across the same gene region. Abbreviations: NN- Nxf1:Nxt1, EV- Empty Vector.

Lastly, Figure 22, Panel D again demonstrates the same region on MUS81, but displays the mapped long read data sequenced from total RNA of the Nxf1:Nxt1 over-expression and total RNA from the EV control. In this case there are examples of long reads clearly spanning the 5' exon through the intron into the 3' exon in both the Nxf1 overexpression sample as well as the EV control. This demonstrates a qualitative example of an IR mRNA present in the nucleus under both experimental conditions, but only present in the cytoplasm with the addition of Nxf1/Nxt1.

For additional examples, Figure 23, Panel A and B demonstrates the cytoplasmic long read data overlying the retained intron of LYPLA2 and LONP1 (both genes with CTEs identified in the vector trap). Both examples demonstrate multiple long reads that span the entire exon-intron complex representing a biologically consistent intron retention event in the cytoplasm.

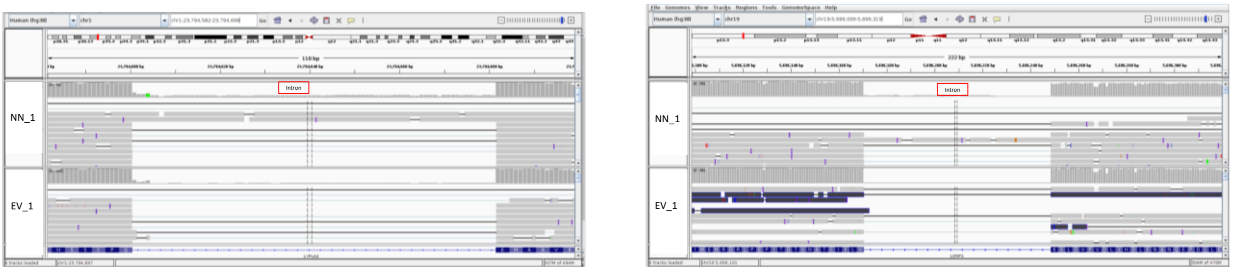

Figure 23: Nxf1:Nxt1 increases cytoplasmic intron retention in numerous cellular genes which contain cellular CTEs. Panel A represents the cytoplasmic long read data from the Nxf1:Nxt1 over-expression experiment described in Figure 22 overlying a retained intron in LYPLA2. Similarly, Panel B represents the cytoplasmic long read data from the Nxf1:Nxt1 experiment overlying LONP1. Abbreviations: NN- Nxf1:Nxt1, EV- Empty Vector.

Discussion

In recent years, it has become clear that intron retention plays an important role in gene expression and regulation in mammalian systems [8, 9, 13, 14]. Despite well described restrictions of nuclear export of mRNA with retained introns [7, 28, 29] as well as the NMD cytoplasmic surveillance system that are assumed to degrade IR mRNAs which contain premature stop codons [31], increasing evidence suggests a subset of retained introns are efficiently exported and translated leading to novel isoform expression and thus increased protein diversity [25, 158, 161, 162]. For example, ROBO3 is a gene involved in axonal guidance during embryonic neural development that retains an intron which is subsequently translated (Intron 26). The IR-protein (Robo3.2) is an antagonistic isoform to the completely spliced protein (Robo3.1), and modulates commissural axons crossing the midline brain [158]. Another example is the ID3 gene where the fully spliced mRNA is expressed following vascular injury, promoting smooth muscle proliferation. However, the ID3 gene also produces an IR-mRNA isoform, that is translated into a protein that has function to limit proliferation [162, 193]. With growing examples of translated IR-isoforms, this indicates that specific mechanisms exist in mammalian cells that allow specific intron containing mRNA to both be efficiently exported from the nucleus, as well as to escape NMD and be efficiently translated.

Previous work on the Nxf1 gene, as well as numerous studies of post-transcriptional regulation in retroviruses, indicate that a cis-acting signal in mRNA paired with an appropriate nuclear export protein, serves to provide export competence for an IR-mRNA [155, 161, 167, 194]. Utilizing the described HIV vector trap, our data demonstrate that this kind of mechanism is likely to be used by many mammalian genes with potentially broad effects on RNA regulation. We specifically demonstrate that cCTEs exist in several previously described mammalian genes that have protein coding IR isoforms (ANKRD1) [176]. We additionally show that cCTEs are globally enriched in cellular pathways that have been previously described as regulated by intron retention, such as neural development and cell differentiation [15]. By identifying and isolating the cCTE sequences from the vector trap and testing them in the dual-color lentiviral reporter assay, our data also demonstrates that the elements have different functional characteristics, including different baseline functionality in 293T cells as well as response to the Nxf1/Nxt1 heterodimer.

The vector trap utilizing the COS cell cDNA library had the advantage of identifying functional RNA elements, as opposed to relying on sequence homology with the original Nxf1-CTE. However, as the serial transfections and hygromycin selection process was time and labor intensive, only a subset of resistant colonies were ultimately identified, and pooled. Furthermore, only a subset of the pooled colonies, were eventually sequenced on the ONT and MiSeq platforms.

Thus, the list of cCTEs identified in this investigation is almost assuredly only a subset of the total number of functional cCTEs in mammalian systems. It was not our intention to discover all possible cCTEs in a mammalian genome, but to determine if CTEs represent a conserved mechanism regulating export of IR-mRNA in multiple mammalian genes. In addition, intron retention is clearly cell type and cell state dependent. Repeating the experimental methodology utilizing, for example, a human neuronal cDNA library would likely generate novel genes containing CTEs and may give insightful information regarding regulation of IR mRNA expression in neuronal cells. Similarly, it would be equally interesting to utilize different cDNA libraries from various disease states, such as undifferentiated human cancers, to discover what elements were unlocked in these circumstances and whether they could affect post-transcriptional regulation in the cancer state, which is a growing field of investigation [195].

Nearly all the cCTEs pulled from the original vector trap experiment mapped to either the 3'UTRs, spanned multiple exons, or were located within introns. There were very few elements that mapped to 5'UTR or single exons. There is previous literature describing the post-transcription role of RNA secondary elements located in the 3'UTR [196]. For elements that had a cCTE span multiple exons, at least partial splicing would have to occur prior to recognition proteins binding to the mRNA. That would infer that only the subset of transcripts that have the exon pattern that form the CTE in combination with the retained intron would be exported, as demonstrated by our MUS81 example. Conversely, cCTEs located in the intron of the gene, would be a simple form of intron self-regulation, as demonstrated by the NXF1 gene [161]. If the intron is present in the mature mRNA, the intron itself carries the secondary structure to allow export and is not dependent on additional splicing patterns.

When we reviewed the genes that contained potential CTEs, several patterns emerged. First, for the cCTE genes that we directly mapped to human protein coding genes, 44 of 55 (80%) had annotated intron containing transcripts in Ensembl. This is consistent with an investigation of 2,500 human mRNA samples demonstrated that IR mRNA isoforms exist in approximately 80% of all protein coding genes [9]. Furthermore, a GO analysis revealed that genes with cCTE overwhelming clustered in cell pathways associated with neuronal development and differentiation. Previous work by Buckley et al demonstrated that intron retention plays a critical role in neuronal biology [15]. The group specifically demonstrated that intron sequences are retained in dendritically targeted mRNA. The group further suggested that a form of SINE retrotransposon on the mRNA may be responsible for dendrite targeting of mRNA. The convergence on biological function suggests cCTEs may play a mechanistic role in the nuclear export privilege of this class of transcripts. Additionally, Pimentel and colleagues demonstrated that differentiating erythroblasts execute a dynamic intron retention program with specific IR events in genes related to RNA processing [12]. They specifically noted a concentration of IR in spliceosome factors, such as SF3B1, and solute carrier genes (SLC) such as SLC25A37 and SLC25A28. In our GO analysis of cCTE genes, RNA binding and processing were similarly over-represented, specifically with numerous genes representing splicing factors and SLC genes.

The results from the Nxf1/Nxt1 over-expression RNA-seq experiment nicely overlapped with many of the findings from the CTE analysis. Nxf1/Nxt1 over-expression increased cytoplasmic intron retention in several hundred protein coding genes over an experimental control utilizing numerous bioinformatic approaches. Interestingly, many of the genes that demonstrated intron retention with Nxf1/Nxt1 over-expression were involved in cellular process of RNA-binding, RNA-splicing and mRNA metabolism. There was also significant direct overlap between genes with CTEs and increased cytoplasmic intron retention with Nxf1/Nxt1 over-expression. A key example discussed in this investigation is MUS81, a DNA endonuclease with essential roles in homologous recombination [197]. The CTE sequenced from MUS81, was functional with the addition of Nxf1:Nxt1 in the supervector experiment. Additionally, the long-read sequencing data from the ONT platform clearly demonstrated full intron containing transcripts in the cytoplasm from this gene. The overall convergence of the CTE and Nxf1/Nxt1 data has two major implications. The first is that many cellular CTEs likely utilize the Nxf1/Nxt1 nuclear export pathway. The second is that cellular post-transcriptional regulation processes, specifically mRNA metabolism (nucleotide binding and splicing) are likely to be self-regulated to some degree by intron retention. Though there is data to support this in the literature, this is one of the first investigation to directly highlight the role of intron retention in post-transcriptional regulatory systems in human molecular systems.

The exact molecular mechanism of nuclear retention of intron containing mRNA itself is not fully understood despite many years of investigation. While it appears that all post-transcriptional pathways of mRNA processing are intimately coordinated including splicing, nuclear localization, nuclear export and cytoplasmic localization, the individual roles of each of these processes in regards to nuclear retention of intron containing mRNA remains unclear. Some investigations support that splicing leaves a 'mark’ on the mRNA, either through the establishment of Exon-Exon Junction Complex or some unknown protein that can recruit nuclear export receptors (like Nxf1) [186, 187, 198-200]. However, the IR-mRNA isoform of Nxf1has multiple introns that are completely spliced 5’ and 3’ to the retained intron 10. Yet, the IR-mRNA is only exported in the presence of the CTE despite having any number of ‘signals’ from the completely spliced exon-exon junctions [161]. It is possible that splicing of all the additional introns occurs co-transcriptionally [201] and the CTE has a functional role at the level of alternative splicing in addition to recruiting the Nxf1/Nxt1 heterodimer. Some evidence also suggests that following splicing, association of paraspeckle proteins may be involved in nuclear retention [202].

In addition to data supporting a role of intron detention at the level of splicing or nuclear trafficking, our group has previously shown that a 'gate-keeping' mechanism may exist at the nuclear basket to restrict nuclear export of intron containing mRNA [203]. More specifically, it was shown that the nuclear basket structural protein Tpr may play a quality control role of mRNA trafficked through the Nxf1 pathway, as even moderate siRNA knockdown of Tpr significantly increased export of IR-mRNA [203]. This is similar to the proposed role of the Mlp1/Mlp2 proteins (Myosin-like Proteins 1 and 2) in budding yeast which also mediates nuclear retention of unspliced mRNAs [204-206]. In addition, our previous work on the Nxf1 gene and the sNxf1 protein suggests that intron retention is tightly regulated and highly tissue specific [25]. For example, the sNxf1 protein is clearly expressed in the hippocampus and cortex cells, but not in other neuronal cells [25]. This finding of IR specificity in both time (cell development/ differentiation) and space (cell type) is supported by additional investigations of intron retention in mammalian systems [10]. In short, it is clear that additional research on the regulatory mechanism of post-transcriptional regulation of intron containing mRNA is necessary.

In this investigation we demonstrated that many mammalian genes are potentially regulated by cCTEs and further identify the specific sequences that can function to medicate export of intron containing mRNA. The data also demonstrates that each element has different levels of functionality in regard to nuclear export capability. Additionally, this data supports that the Nxf1/Nxt1 export pathway continues to be a critical pathway for the export of IR mRNA through the CTE regulated mechanism and may regulated numerous genes in this manner. Attention has recently turned to specific mRNA architecture and sequence composition of mRNA that may be directly regulated by Nxf1, as opposed to through adapter proteins as is usually described for bulk mRNA export of the Nxf1 pathway [189]. The direct correlation of Nxf1 expression to nuclear export and translation of IR-mRNA, as well as differential transcript expression of protein coding genes is another area of current investigation of our team.

Conclusion

The data from the study supports that cis-acting regulatory structures on RNA, termed CTEs, are functionally active domains that allow IR-mRNA to gain nuclear export privilege in mammalian systems. We demonstrate that functional cellular CTEs exist in hundreds of mammalian genes with different levels of functionality and with different levels of responsiveness to Nxf1:Nxt1 expression. The genes that are potentially regulated by CTEs appear to cluster in cellular pathways involved in cell differentiation and development, specifically for neurons. At a molecular level, the genes seem to be involved directly in post-transcriptional regulation including specifically mRNA binding. Our data further supports that previously described genes with translated intron retention, such as ANKRD1, have functional CTEs which are likely responsible for the export and the translation privilege of IR-mRNA isoforms.

Methods

Plasmids and Cloning Procedure

All plasmids utilized in this chapter follow the nomenclature pHRXXXX for identification and ease of requests. For the vector trap experiment, the backbone plasmid of the modified NL4-3 HIV provirus with intact LTRs and GagPol message with the internal SV40 driven hygromycin resistence and an empty stuffer fragment is denoted pHR2016. Numerous individual plasmids were generated from this backbone by cloning various COS cDNA fragments in to the BstxI restriction sites of the pHR2016 backbone. Individual colony generation plasmids take the specific plasmid format- pBEB-NL43-F12IS-clone#. Following identification of the cellular CTEs in the vector trap, specific COS cDNA fragments (cCTE) were cloned into the pCMV-GagPol (pHR2739) reporter plasmids lacking an RRE as previously described [27]. The plasmid is based on the pCMV-GagPol-RRE (pHR0354) plasmid that was originally used to demonstrate Rev function [207]. The plasmid and cloning process positions the cellular insert immediately downstream of the of the intron containing GagPol message and in a position to act as a functional RNA export element for the intron containing GagPol message.

The dual color reporter vector (Supervector) experiment is based on the original pNL4-3(eGFP)(NL4-3 RRE)(mCherry) (pHR5604) vector described previously [173]. The Supervector was specifically designed to easily substitute different nuclear export elements in place of the RRE element. It was also designed to have different nuclear export proteins provided *in-trans*, setting up nicely to screen the functionality of different combinations of RNA nuclear export elements and nuclear export proteins. Of note, we did create a VPR- Supervector backbone (pHR5753) and used this vector for our cellular CTE inserts as VPR can be toxic to cells. To compare results with our previously unpublished data we inserted the NXF1-CTE (pHR5753), the ACTN4-CTE (pHR6336), and the SIRT7-CTE (pHR6332) into the Supervector backbone. We utilized the RRE Supervector (pHR5604) without the addition of Rev as a negative control. We used pCMVNxf1 (pHR3704) and pCMVNxt1 (pHR2415) for our *trans*-acting nuclear export proteins in all Supervector experiments. Following this experiment, we screened over 20 cellular CTE elements by the same process, inserting the respective cCTE in place of the RRE. Key plasmids tested in this chapter included ACTN1-CTE (pHR5612), ANKRD1-CTE (pHR5614), RPL18A-CTE (pHR5616), SMARCB1-CTE (pHR5618), Nxf1-CTE (pHR5753), XRN2-CTE (pHR5860), PYCR2-CTE (pHR5876), MPMV-CTE (pHR5886), RPLP0-CTE (pHR6064), EIF3G-CTE (pHR6066), COL4A2-CTE (pHR6134), HSP90B1-CTE (pHR6135), LONP1-CTE (pHR6136), MUS81-CTE (pHR6137), TCEAL8-CTE (pHR6139), PRRC2-CTE (pHR6138), CCT6A-CTE (pHR6140), TSR2-CTE (pHR6142), CRKL-CTE (pHR6148) and NONO-CTE (pHR6199), TMTC3-CTE (pHR6467), 5A1-CTE (pHR6334), and 6A29(MXRA7)-CTE (pHR6330). For the Nxf1:Nx1 overexpression experiment 3 plasmid vectors were utilized. pCMV-Empty (pHR0016), pCMV-NXF1 (pHR3704) and pCMV-NXT1 (pHR2415).

The major cloning procedure used in this Chapter for the Supervector constructs was the NEBuilder HiFi DNA assembly kit (New England Biolabs, Inc). The cellular CTE genes blocks were designed with a ~25 bp overlap with the linearized Supervector plasmid backbone- pNL4-3(eGFP)(NL4-3 RRE)(mCherry) (pHR5604). The Supervector backbone was cloned with individual cellular CTE inserts in a 1:4 ratio as suggested by the manufacturer's instructions. Following assembly, the cloned plasmid sequence (sent with primers) was confirmed using standard Sanger DNA sequencing (Eton Bioscience Inc, San Diego, CA).

Cell Lines and Cell Transfections

For the Vector trap experiment, B4.14 packaging cell line was maintained in Dulbecco's Modified Eagle Medium (DMEM) with 10% fetal calf serum, 50 mg/ml gentamicin and 200 mg/ml hygromycin B, and Hela cells were maintained in DMEM with 10% fetal calf serum and 50 mg/ml of gentamicin. The B4.14 packaging cell line was transiently transfected using a Calcium Phosphate protocol previously described [26, 208]. 293T cells were maintained in DMEM, 10% Bovine Calf Serum (BCS) and 50 mg/ml gentamicin. Transient transfections were performed using a lipofectamine3000 protocol in 293T Cells as suggest by the manufacturer's instructions (Thermofisher Scientific).

Vector Trap Experiment and DNA Sequencing

The B4.14 packaging cell line, which constitutively produces HIV-1 GagPol proteins, was transfected with vectors containing either the RRE, CTE, no RRE/CTE (DRRE) or the vector library containing COS cDNA fragments. Three days post-transfection, supernatants were collected and used to infect Hela cells, which were subsequently selected in hygromycin-containing medium for 14 days. Resistant colonies were then subjected to a second round of vector mobilization by cotransfecting plasmids expressing HIV-1 GagPol and VSVG envelope proteins to rescue the integrated HIV vector sequence from the genome. Supernatants from transfected cells were used to infect unaffected Hela cells that were again selected in hygromycin-containing medium. Genomic DNA was isolated from resistant colonies and used as a template to amplify cDNA fragments using PCR primers flanking the cloning sites in the vector. COS cDNA library fragments, which contained cellular CTEs, were then sequenced on MiSeq and Oxford Nanopore Long Read Sequencing Platforms and mapped to the human genome (please see below for sequencing and mapping details).

Sequencing and Genomic Alignment of Cellular CTEs

We designed specific DNA primers along the flanking restriction sites along the vector backbone (restriction sites 5' Mam1 and 3' Hpa1) which defined the stuffer fragment and cellular insert from the COS cell cDNA library. These primers were utilized as PCR primers for both the MiSeq and ONT sequencing. Isolated DNA from pooled colonies were subjected to specific PCR amplification protocols for both sequencing platforms. For the ONT sequencing we used a cDNA PCR amplification library preparation protocol (SQK-LSK109) as described by the manufacturer's protocol. The cDNA library was then sequenced on a Flongle using the MinIT. Raw ONT signals were processed using MinKNOW Version 2.2 to generate FASTQ files. FASTQ files were then trimmed using a custom CutAdapat script to remove stuffer fragments to the flanking restriction sites of the cDNA insert itself (flanking restriction sites 5' BstX1 and 3' BstX1). Trimmed FASTQ files were then mapped to the human genome (Hg38) using BBMAP. Mapped reads were grouped by genes and imported in Geneious (Biomatters, Auckland, New Zealand) for visual inspection and to generate a consequence sequence of the cDNA insert. Of note, the shortest functional sequence that aligned across the gene region was used to defined the minimum cellular CTE sequence for functional testing. For the MiSeq platform, we similarly performed a single round of PCR using our specific primers to isolate the cellular fragments from hygromycin resistant colonies. We then used a MiSeq Reagent Kit v2 Nano for library preparation and sequenced on a MiSeq machine.

Supervector Assay and Flow Cytometry Parameters and Statistical Analysis

The Supervector assay and flow cytometry assay was performed according to the method publication specifications [173]. For the specific experiments in this chapter, 293T cells ($4.5x10^6$ cells in 6 well plate, 2 ml of media) were transfected with 1,000 ng of the Supervector plasmid with the *cis*-acting cellular CTE insert with and without 100 ng of pCMVNxf1 and 200 ng of pCMVNxt1. Cells were harvested 48 hours post-transfection and prepared for flow-cytometry analysis. As described in the publication, flow cytometry was performed on a Attune NxT flow cytometer with an attachment autosampler (Thermofisher Scientific). Data analysis was performed using FlowJo v10 (FlowJo, LLC). The gating strategies were performed as suggested in the publication- live cells were selected first (forward-scatter vs. side-scatter) followed by single cell selection (forward-scatter area vs forward-scatter height). The mean fluorescent intensity (MFI) of eGFP and mCherry for each individual cell was recorded. Functional activity of a particular CTE was determined by the ratio of eGFP to mCherry. Each experimental condition was always performed in triplicate

Numerous independent transfection experiments were conducted to test all the various cellular CTEs at different time points. Each transfection experiment was conducted with the same three experimental controls- Supervector with RRE (no Rev), Supervector with Nxf1-CTE (with and without Nxf1:Nxt1) and Supervector with MPMV-CTE (with and without Nxf1:Nxt1). To compare cellular CTE functionality across independent transfections conditions, each transfection experiment was normalized based on the averaged ratio of these three experimental controls. The geometric mean and standard deviation of the eGFP/mCherry ratio for each experimental condition was calculated in R (version 3.5.1). Statistical hypothesis testing was conducted using a two-sided Student’s t-test. Statistical significance was determined by p value < 0.05 and a 1.5-fold change above the comparison value.

Nxf1:Nxt1 Over-expression Experiment and NGS Sequencing

293T cells (10 cm plates, $3x10^6$, 10 ml media) were transfected with 5 ug of pCMVNxf1 and 10 ug of pCMVNxt1 or with 15 ug of pCMV-Empty-Vector Insert. Cells were harvested 72 hours post-transfection. Each experimental condition was performed in triplicate in two separate experiments (6 total transfected plates per experimental condition). 3 cell plates per transfection condition were utilized for either cytoplasmic or total RNA isolation for later statistical comparison. A phenol: chloroform extraction protocol was utilized for both total and cytoplasmic RNA isolation as previously described [209]. Key steps to the cytoplasmic protocol include washing pelleted cells with ice-cold phosphate-buffered saline (PBS) followed by resuspension in a reticulocyte standard buffer (RSB) (10 mM Tris-HCl (pH 7.4), 10 mM NaCl, 1.5 mM MgCl2) followed by the addition of RSB with 0.1% IgePal in equal volume to perform cell lysis. Cell nuclei were pelleted out using two sequential centrifugations at 13,000 rpm in an Eppendorf centrifuge at +4C. 2 PK buffer (200 mM Tris-HCl (pH 7.5), 25 mM EDTA, 300 mM NaCl, 2% SDS, 400 ug of proteinase K/ml) was added to cell lysates (with nuclei cleared) and incubated at 37C for 30 min. Cytoplasmic RNA was then extracted from the lysate using a phenol and chloroform-isoamyl alcohol (24:1) in a 1:1 volume followed by 2 separate extractions with chloroform-isoamyl alcohol (24:1). RNA was precipitated using 200 proof ethanol (-80C) at a 2.5 volume with the addition of sodium acetate (3M, pH 5.5- final concentration 0.3 M. RNA precipitated in EtOH was stored at -80C. Key steps for the total RNA extraction including washing pelleted cells twice with ice cold PBS. Cells were lysed with lysis buffer containing 0.2M Tris-HCL (pH 7.5), 0.2 M NaCl, 1.5 mM $MgCl_2$, 2% sodium docdecyl sulfate (SDS) and 200 ug of proteinase K per ml and incubated for 1 hour at 45C.

RNA-EtOH slurry was pelleted for 1 hour at 4C at 3,000 RPM (3014g) in a Baxter Cryofuge 6000 centrifuge. RNA pellet was washed twice with 75% ethanol and air dried twice. Total RNA additionally DNase digested (RQ1 RNase-free DNase, Promega) then extracted as above. RNA was resuspended in 10 ul of RNase/DNase free water. RNA concentrations were determined by Qubit RNA HS assay kit (Qubit 3.0 fluorometer, Thermofisher). RNA quality was determined by RNA Tape station. PolyA+ RNA was isolated from the cytoplasmic RNA using poly-T magnetic bead-based protocol according to the manufactures protocol (NEXTFLEX Poly(A) Beads 2.0 Kit, PerkinElmer Inc).

Isolated RNA was subjected to both short read and ONT long reads sequencing. Short read sequencing was performed at Novogene (Beijing, China). Isolated polyA+ RNA was shipped to Novogene (on dry ice) who reconfirmed RNA quality and prepared stranded cDNA libraries. Libraries were sequenced on an Illumina HiSeq3000 system to generate paired-end 150 bp reads to a sequencing depth of approximately 40M reads per sample. From each transfection condition (cytoplasmic Nxf1:Nxt1 RNA, total Nxf1:Nxt1 RNA, cytoplasmic EV RNA, total EV RNA) equimolar amounts of RNA was isolated from each triplicate experiment and the RNA pooled for a total of 50 ug of polyA selected RNA for each condition. A direct cDNA native barcoding ONT protocol (#SQK-DCS109) was used to prepare our cDNA library from the pooled polyA+ RNA following the manufactures instructions (Oxford Nanopore Technologies, United Kingdom). Following barcode and adapter ligation the cDNA library was loaded on a single SpotON Flow cell with the appropriate amount of sequencing buffer and loading beads. The sequencing was performed with the assistance of the MinIT device (active Base-calling ON) which processed raw signals and converted them to nucleotides (FASTQ format) via MinKNOW v2.2 program.

A direct cDNA native barcoding ONT protocol (#SQK-DCS109) was used to prepare the cDNA library from the polyA+ cytoplasmic RNA following the manufactures instructions (Oxford Nanopore Technologies, UK). Following barcode and adapter ligation, the cDNA library was loaded on a single SpotON flow cell with the appropriate amount of sequencing buffer and loading beads. The sequencing was performed with the assistance of the MinIT device (active Base-calling

ON) which processed raw signals and converted them to nucleotides (FASTQ format) via MinKNOW v2.2 program.

Bioinformatic Analysis for Intron Retention Discovery of short read RNA-seq data

Four separate bioinformatic techniques were utilized for intron retention discovery. For each pipeline, Nxf1:Nxt1 overexpression was compared to EV control (each condition had biological triplicates). The splice aware differential gene expression pipeline- HiSAT2, Stringtie and Ballgown with recommendation conditions (genome and annotation file used were Ensembl Human Genome Hg38.v37 and corresponding GTF file) was used first [154]. Similar to DEXSeq, Ballgown allows quantification across specific gene-level features including at the transcript, exon and intron level. The intron genomic position is determined by the GTF file used during read assembly step (Stringtie). This allowed both quantify and calculation of a differential expression across all reads that aligned within intronic positions between conditions. Note, this analysis did not take into account splice junction reads, it simply included a normalized count measure for any read within the mapped intron. Secondly, a differential transcript expression analysis using Salmon and DESeq2 was performed according to the suggested parameters (Ensembl GRCh38.95 cDNA FASTA File) [210]. A custom R script was used to pull all transcripts defined as having a retained intron from the GTF file from Ensembl (GRCH38.95). This transcript list was used to sub-select all differential transcripts in our Nxf1:Nxt1 to EV comparison that contained a retained intron.

rMATS (Multivariate Analysis of Transcript Splicing) was also used to define a percentage spliced in (PSI) ratio of intron retention according to standard recommendations [192]. rMATS calculates several different splicing events including Skipped Exon (SE), Alternative 5’ Splice Site (A5SS), Alternative 3’ Splice Site (A3SS), Mutually exclusive exons (MXE) and Retained Intron (RI). The RI.MATS.JCEC output from the analysis was used to identify differentially expressed retain intron events between conditions. Lastly, IRFinder was used for intron retention discovery and differential analysis according to specifications [9]. IRFinder is based on the STAR alignment algorithm (version STAR-2.7.3a). RNA-seq reads were aligned to Ensembl Human Genome Hg38.v37 and corresponding GTF file. IRFinder calculates an IRatio across different introns. Differential expression of individual IRatio between conditions was performed using the IRFinder provided DESeq2 R script under recommend specifications.

Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) over-representation functional analysis was performed on Genes that contained intron retention events. Gene pathways were considered enriched if they had a p-adj value < 0.05. GO and KEGG analysis was performed using the ShinyGO package for visualization and the clusterProfiler package in R [151, 174].

Bioinformatic Analysis for Long Read ONT data

FASTQ reads were processed with the ONT Pinfish pipeline for long read processing including sub-selecting full length transcripts using the Pychopper algorithm [122]. The sub-selected full-length transcripts were then mapped directly to human genome (Ensembl Hg38.v37) using Minimap2 (using map-ONT mode) [123]. The full length, mapped reads (.BAM files) were sorted by chromosomal position and loaded into IGViewer for visualization of full length reads.

V. The Molecular Role of WT1(+KTS) in Post-Transcriptional Regulation and Intron Retention

Abbreviations

Wilms’ tumor 1 gene- WT1; KTS- amino acids Lysine, Threonine, Serine; mesenchymal to epithelial transition MET; Constitutive Transport Element- CTE

Introduction

Wilms' tumor is an embryonal tumor which develops from undifferentiated nephrogenic rest cells that subsequently undergo malignant transformation [56, 57]. Evidence that the tumors arise from pluripotent malignant embryonal cells includes the typical triphasic histology of Wilms' tumor. The disease is histologically categorized by the International Society of Paediatric Oncology (SIOP) based on the differentiation status of the tumor- low risk tumors are partially differentiated, while high risk tumors have blastemal components or diffuse poorly differentiated anaplastic cells [57]. As a result, research into the biological drivers of these fetal tumors overlap with the fetal development of the kidney as well as failure of organ differentiation.

During fetal development, the kidney arises jointly from the ureteric bud, which forms the collecting duct system, and the metanephric mesenchyme, which forms the nephron [211]. The development of the nephron is dependent on a mesenchymal to epithelial transition (MET), which is regulated by the WT1 gene [56, 211]. Inactivating mutations in both copies of WT1 were found in Wilms' tumor, leading the gene to be classified as one of the original tumor suppressor genes [212, 213]. Mutations in the WT1 gene are also associated with multiple genetic syndromes, which can lead to the development of Wilms' tumor. These include Denys-Drash, Frazier Syndrome and WAGR syndrome [214-216].

Several decades of study have revealed WT1 to be a much more complex gene than initially anticipated. WT1, located at 11p13, has 10 exons and can produce at least 36 potential isoforms [217]. One of the key functional domains of WT1 include four zinc fingers at the carboxy terminus of the proteins, which are structurally similar to the zinc finger found in the SP1 family of transcription factors [59]. The functional importance of all WT1 isoforms is not clear. However, two isoforms, denoted WT1 +KTS and WT1 -KTS, which differ by a 3 amino acid insertion (KTS) between zinc fingers 3 and 4 at the end of the 9th exon, have critical functional consequences.

The WT1 -KTS isoform is an effective DNA binder and is involved in regulating transcription. Numerous investigations have identified genes that are regulated by -KTS and the downstream effects of inactivating mutations of WT1 on transcription [218]. In contrast, the addition of the +KTS sequence in the protein changes the orientation of the 4th zinc finger, significantly reducing DNA binding, but retaining strong RNA binding capacity. WT1+KTS and -KTS are both expressed in the developing kidney, and the ratio of these isoforms (usually 2:1) is critical for proper nephron differentiation. Alterations in the isoform ratio as seen in Frasier syndrome, where a specific mutation in WT1 reduces the use of the 5' splice site that results in the +KTS isoform and reverses the isoform ratio to 1:2, is associated with a high incidence of Wilms' tumor [215, 216]. Despite the significant phenotypic effects that are thus caused by a reduction of the WT1 +KTS isoform, how this protein functions in post-transcriptional gene regulation remains poorly understood.

As discussed in the introduction to this thesis, the Hammarskjold/Rekosh Lab previously demonstrated that the WT1+KTS isoform effectively promotes translation of a model mRNA containing a retained intron [219]. The Hammarskjold/Rekosh Lab further showed that this function was dependent on a constitutive transport element (CTE). As described in Chapter 2 of this thesis, CTEs are cis-acting RNA elements, present in both cellular and viral genes that allow export and expression of mRNAs with retained introns [20, 167]. Using the previously described HIV GagPol reporter vector, the data showed that WT1+KTS isoform increased protein expression from an HIV mRNA with a retained intron by ~30-fold compared to controls. The data further suggested that the WT1+KTS isoform did not affect export, but promoted the polyribosome association of the intron containing mRNA, leading to an increase in protein synthesis [27]. Similar results were achieved when the ACTN4-CTE was used in place of the viral MPMV-CTE [219]. These results suggest that WT1+KTS promotes translation of CTE-containing mRNAs by promoting the association with ribosomes.

The results of these studies led the us to hypothesize that WT1+KTS may serve to specifically regulate expression of alternatively spliced cellular mRNAs, including mRNA with retained introns that contain cellular CTEs. In an investigation to discover which cellular mRNAs were affected by the WT1+KTS isoform (in combination with Nxf1/Nxt1), preliminary experiments were performed in the HamRek lab many years ago. 293T cells were specifically used as they do not produce endogenous WT1 protein [220-222]. Following transfection with plasmids expressing WT1+KTS and Nxf1:Nxt1, polyribosomal fractions were isolated from sucrose gradients. Affymetrix microarray analysis on the mRNA isolated from polyribosomes indicated that multiple genes were differentially expressed compared to mock transfected cells. Several genes were shown to be upregulated greater than 2-fold including EGFR (10-fold), EGR1 (2.7fold), ACTN4 (3.7-fold), as well as multiple heat shock proteins including HSPA6 (34-fold) and HSPA1A (16-fold). However, it was not possible to analyze which isoforms were upregulated using this kind of analysis, which limited the utility of this approach. Although this study was never completed, these preliminary data supported the notion that WT1+KTS promoted polyribosomal association of specific cellular mRNAs.

Most of the post-transcriptional mRNA targets for the WT1+KTS isoform have remained unknown and under-investigated [223, 224], despite the clear biological importance of this protein in health and disease. The lack of identified cellular targets limits both the ability to further investigate what role post-transcriptional regulation may have on cell differentiation, as well as the effect on disease states such as Wilms' tumor. The focus of the work presented in this chapter was to build on previous experiments and specifically identify how WT1+KTS qualitatively and quantitatively affected differentially expressed mRNAs including mRNA with retained introns in cytoplasmic mRNA.

Results

WT1+KTS expression alters expression of many mRNA isoforms in the cytoplasm from genes involved in post-transcriptional regulation

In order to determine which cellular genes and in-turn which cellular processes might be regulated by the post-transcriptional isoform of WT1, I expressed WT1+KTS in 293T cells and performed an RNA-seq based experiment. As discussed in the introduction to this chapter, 293T cells have very minimal endogenous WT1 expression [220-222], making this cell line well suited for these experiments. A WT1+KTS plasmid as well as an Empty Vector control plasmid were transfected in 293T cells using a Lipofectamine3000 protocol. The experiment was performed in triplicate for all conditions in this chapter. After 72 hours, cytoplasmic and total RNA was isolated from cell plates using a Phenol: Chloroform based protocol for each biological replicate. Following precipitation, RNA was polyA selected using a magnetic bead-based protocol. Isolated cytoplasmic mRNA was then sent to Novogene for RNA-sequencing on the Illumina HiSeq Platform. The experiment generated approximately 40M stranded, pair-end reads per replicate. RNA-seq reads were trimmed and quality control checked prior to bioinformatic analysis.

A standard differential gene expression analysis comparing cytoplasmic WT1+KTS to the cytoplasmic EV control was conducted. Reads were aligned to the human genome (Ensembl

Hg38.v37) using HiSat2 and assembled using Stringtie (Ensembl Hg38.v37 GTF). Samples reads were normalized and a differential gene expression analysis calculated using the R program Bioconductor package Ballgown according to the New Tuxedo pipeline [154]. Significance was determined to represent a 1.5 > |log2fc| and an adj p-value < 0.05. The analysis resulted in only 1 gene that was significantly upregulated at the overall gene level, and 3 that were downregulated. Interesting, the upregulated gene was ANKRD1, which is known to have a CTE as discussed in Chapter 2 of this thesis, which demonstrated a 17.6-fold increased expression in the presence of WT1+KTS. However, the predominate finding was that WT1+KTS had very little effect on differential gene expression (DGE) in 293T cells. This is consistent with previous investigations which have suggested that WT1+KTS shows poor binding to DNA and in contrast to WT1-KTS is not a significant transcriptional regulator [225-227].

I next performed a differential transcript expression and differential transcript usage pipeline (which measures different isoform ratios across conditions), comparing cytoplasmic WT1+KTS to EV control utilizing the DRIM-Seq pipeline with suggested parameters from published methodology [210, 228]. I used the pseudoaligner, Salmon [147] in mapping-based mode with the validateMappings flag to create a count matrix over the full human transcriptome (Ensembl GRCh38.95). The count matrix from each biological replicate was imported into R using tximport [148]. A differential transcript expression analysis was first performed using DESeq2. In contrast to the HiSAT2-Stringtie differential gene expression analysis, which found very few changes at the gene level, there were 2,137 transcripts that were differentially expressed using the Salmon pipeline including 1,548 that were upregulated as a result of WT1+KTS expression (Appendix Table 7). Interestingly, one of the known protein coding mRNA isoforms from the proto-oncogene c-MYC (ENST00000621592) demonstrated the highest increased log2fc expression level of any transcript (log2fc = 8.2, p-adj < 0.0001). I performed a formal Gene Otology (GO) Enrichment Analysis of the differentially expressed transcripts (Figure 24). WT1+KTS expression affected several specific cellular processes, including clear effects on genes involved in post-transcriptional regulation.

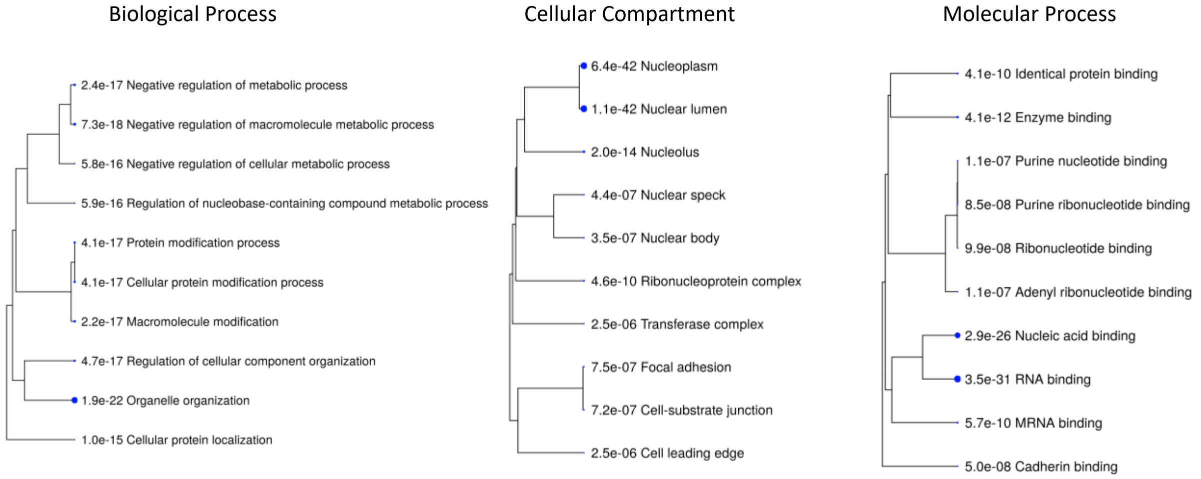
Biological Process
2.4e-17 Negative regulation of metabolic process
7.3e-18 Negative regulation of macromolecule metabolic process
5.8e-16 Negative regulation of cellular metabolic process
5.9e-16 Regulation of nucleobase-containing compound metabolic process
4.1e-17 Protein modification process
4.1e-17 Cellular protein modification process
2.2e-17 Macromolecule modification
4.7e-17 Regulation of cellular component organization
1.9e-22 Organelle organization
1.0e-15 Cellular protein localization
Cellular Compartment
6.4e-42 Nucleoplasm
1.1e-42 Nuclear lumen
2.0e-14 Nucleolus
4.4e-07 Nuclear speck
3.5e-07 Nuclear body
4.6e-10 Ribonucleoprotein complex
2.5e-06 Transferase complex
7.5e-07 Focal adhesion
7.2e-07 Cell-substrate junction
2.5e-06 Cell leading edge
Molecular Process
4.1e-10 Identical protein binding
4.1e-12 Enzyme binding
1.1e-07 Purine nucleotide binding
8.5e-08 Purine ribonucleotide binding
9.9e-08 Ribonucleotide binding
1.1e-07 Adenyl ribonucleotide binding
2.9e-26 Nucleic acid binding
3.5e-31 RNA binding
5.7e-10 MRNA binding
5.0e-08 Cadherin binding

Figure 24- WT1+KTS regulates cellular organization and macromolecule metabolic processes including nucleic acid and protein modification. 293T cells (10 cm plates, 3x10^6, 10 ml media) were transfected with 15 ug of WT1(+KTS) or pCMV-Empty-Vector. Cells were harvested 72 hours post-transfection. Each experimental condition was performed in triplicate. Cytoplasmic mRNA was isolated using a Phenol: Chloroform protocol and sequenced on an Illumina HiSeq platform (stranded 150 bp paired-end reads). A differential transcript expression analysis was run using Salmon. Significantly expressed transcripts (|log fold change| < 1.5, p-adj < 0.05) were used for a Gene Ontology Enrichment Analysis performed using the ShinyGO visualization platform. Gene pathways that demonstrated an enrichment false discovery rate < 0.05 were included. Displayed above are the top 10 gene pathways (organized by FDR significance) in each respective category of the GO analysis- Biological Process, Cellular Component and Molecular Process.

The differential transcript usage analysis demonstrated that the WT1+KTS changes specific mRNA isoforms in the cytoplasm. There were 1,313 genes that showed significant changes in isoform ratios with WT1+KTS expression (padj value < 0.05), which included 1,967 transcript specific isoforms (Appendix Table 8). Interestingly, c-MYC demonstrated a near complete isoform switch between two different protein coding isoforms in the cytoplasm of 293T cells following WT1+KTS expression (Figure 25, Panel A). The expression of WT1+KTS caused an isoform switch from the predominate c-Myc2 isoform (a 3-exon isoform which utilizes a canonical AUG start codon following a fairly long 5' untranslated exon) to c-Myc1 isoform, which is a 3-exon isoform which utilizes a non-canonical CUG start codon 15 codons upstream of the c-Myc2 start site. These two isoforms produce 2 proteins with distinct functions [229]. Whereas c-Myc2 isoform predominately mediates transcriptional activation and is upregulated during cell growth, c-Myc1 appears to arrest growth and leads to apoptosis [229, 230].

Additionally, WT1+KTS expression appeared to induce specific isoforms in the cytoplasm that were not present under the control conditions as in the case of the EIF5 (Figure 25, Panel B). The newly expressed isoform, EIF5-214, is a protein coding isoform and appeared to decrease the isoform ratio of the other 4 expressed EIF5 transcript isoforms. The ribosomal protein, RPL36 demonstrated ratio changes in 2 protein coding isoforms with WT1+KTS expression, though both isoforms were still maintained in the cytoplasm (Figure 25, Panel C). In the case of ACTB, the newly expressed isoform ACTB-208 is a protein coding isoform and predominated over an ACTB processed transcript (Figure 25, Panel D).

I repeated the differential transcript usage pipeline using the DRIMSeq methodology under identical parameters for total WT1+KTS RNAseq data compared to total EV control. There were 407 genes that demonstrated significant isoform ratio changes with WT1+KTS expression (Appendix Table 9). Interestingly, the genes demonstrating isoform ratio changes in total RNA had a 142 gene overlap with the genes demonstrating an isoform ratio change in the cytoplasm following WT1+KTS expression (10.8%). For example, c-Myc did not demonstrate an isoform ratio difference in the total RNA data as it did in the cytoplasmic data. There was also clearly a reduced effect on differential isoform ratios in the total data (400 effected genes) as in the cytoplasm (1,300 effected genes) following WT1+KTS expression. The RNA-seq data presented does not address the molecular mechanism by which WT1+KTS preferentially affects transcript isoform ratios, though the data does support that WT1+KTS has an important role in cytoplasmic utilization of mRNA. However, given that there were isoform ratio changes in genes in the total RNA that were not present in the cytoplasmic RNA data, WT1+KTS clearly has a complex post-transcriptional role with possible effects on nuclear RNA processes (possibly splicing) in addition to cytoplasmic RNA processes (RNA stabilization or ribosomal association of mRNA).

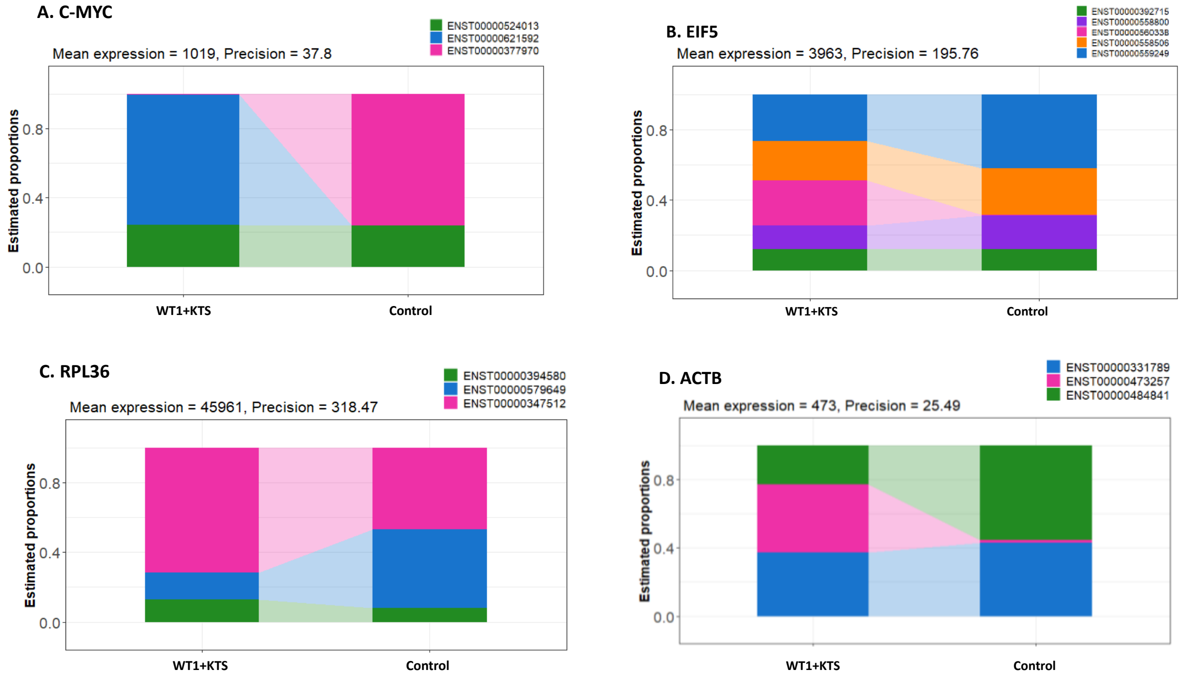

Figure 25, Panel A: WT1+KTS displays significant transcript specificity in a differential transcript usage analysis. RNA-sequencing reads from WT1+KTS and EV control were subjected to a differential transcript usage pipeline utilizing Salmon pseudoalignment and quantification with DRIMseq. Significance was determined by a |fold change| > 1.5 and p-adj value < 0.05. The figure represents the estimated transcript expression proportions across various genes including Panel A. c-MYC, Panel B. EIF5, Panel C. RPL36, Panel D. ACTB.

Many of the genes found to be altered by WT1 +KTS expression were previously shown to express mRNAs that bound directly to the WT1 protein in a previously published study

I next wanted to determine which of the cytoplasmic changes that were observed after WT1+KTS expression could be a result of direct mRNA binding. Hastie and colleagues previously performed an RNA immunoprecipitation (RIP) and sequencing (RIP-seq) experiment to identify WT1 interacting RNAs in mouse embryonic stem cells [196]. The group discover ~2,500 RNA binding targets in the embryonic stem cell line which was consolidated at the gene level. When we compared these to the 1,820 genes that had differential transcript expression with WT1+KTS expression in the cytoplasm, we noted a 390 gene overlap which was statistically significant (Fisher's exact test, $p < 0.0001$). We performed a GO enrichment analysis on the 390 gene overlap and found similar cellular and molecular processes affects as the original DTE analysis (Figure 26). Specifically, the biological processes regulated by direct RNA binding targets following WT1+KTS expression include RNA splicing, nucleocytoplasmic transportation and protein binding.

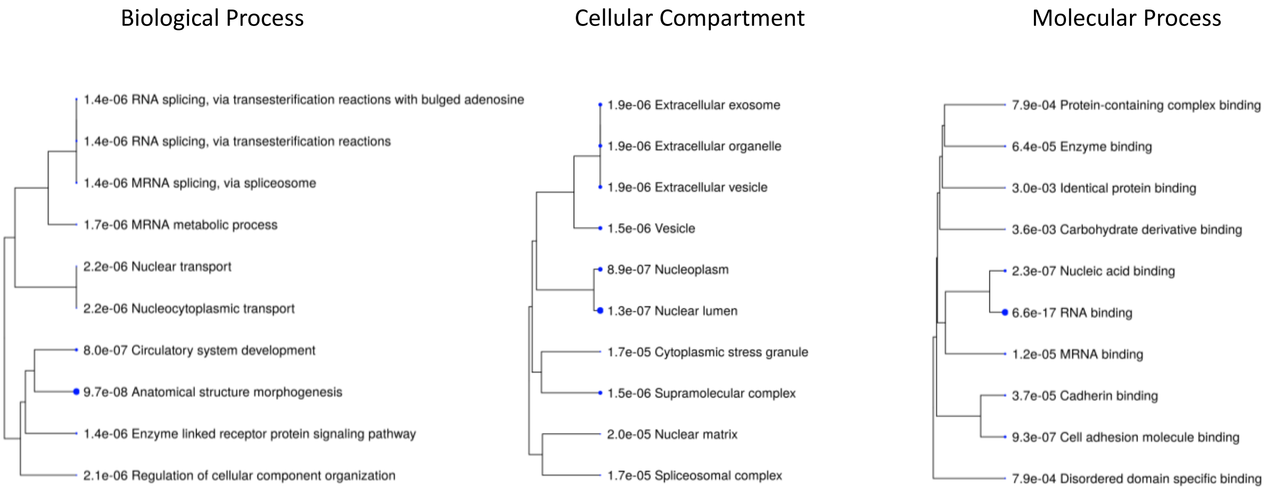

Figure 26: WT1+KTS post-transcriptional regulation is a result of direct mRNA binding to transcripts involved in splicing, nucleocytoplasmic trafficking and protein binding. The list of differentially expressed transcripts following WT1+KTS expression was compared to a list of WT1 direct RNA binding targets from an immunoprecipitation pipeline (Hastie et al. 2017). Genes with direct overlap were used for a Gene Ontology Enrichment Analysis performed using the ShinyGO visualization platform. Gene pathways that demonstrated an enrichment false discovery rate < 0.05 were included. Displayed above are the top 10 gene pathways (organized by FDR significance) in each respective category of the GO analysis- Biological Process, Cellular Component and Molecular Process.

WT1 +KTS increases mRNA with retained introns in the cytoplasm in genes shown to be direct RNA binding targets of WT1

WT1 +KTS works synergistically with Nxf1:Nxt1 to increase protein production from an mRNA with a retained intron containing the viral MPMV-CTE

As briefly discussed in the introduction, the Hammarskjold/Rekosh lab previously demonstrated that the WT1+KTS protein isoform interacts with the various cellular CTEs to promote polyribosome association and ultimately increased protein formation of intron containing mRNA [27]. To redemonstrate the WT1(+KTS) increases translation of mRNA with a retained intron in 293T cells, I utilized a previously described HIV reporter construct containing the Gag-Pol intron (Figure 27 Panel A) to measure the translation efficiency of IR-mRNA utilizing a p24 ELISA assay under various conditions with WT1 +KTS. The reporter produces an unspliced mRNA with a retained intron that encodes the HIV Gag and GagPol proteins. In the absence of a functional export element, the mRNA is retained in the nucleus. However, if an RRE is inserted into the vector and Rev is supplied in *trans*, the mRNA that is produced reaches the cytoplasm and is efficiently translated into protein [20, 26, 161, 167]. Our lab has also shown that both the MPMV-CTE and the Nxf1-CTE can be inserted in place of the RRE in the HIV-Gag-Pol vector to mediate nuclear export and translation. The levels of p24 increase considerably when exogenous Nxf1 and Nxt1 are added [161].

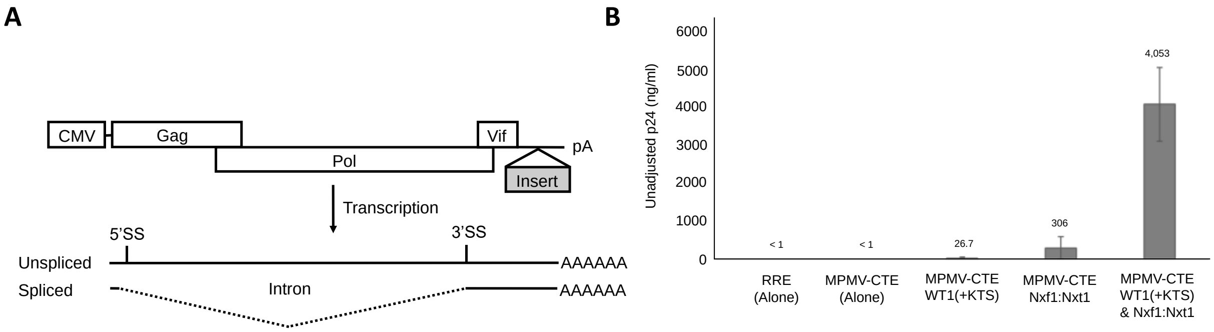

Figure 27: WT1 +KTS works synergistically with Nxf1:Nxt1 to increase protein production from an mRNA with a retained intron containing the viral MPMV-CTE. Panel A: Schematic representation of reporter constructs used in this experiment. The Insert can be any export element- RRE, MPMV-CTE, Nxf1-CTE, or other cellular CTEs. In the presence of a functional cCTE, the unspliced mRNA is exported to the cytoplasm. The dotted line below the spliced mRNA depicts the location of an intron. For this experiment, the insert was replaced with either the HIV RRE and the MPMV-CTE. Panel B: 293T cells (5.5x10^5 cells in 6-well plate, 2 ml media/well) were transfected with 1,000 ng of pCMV-GagPol reporter plasmid with either the RRE or the MPMV-CTE in addition to 1,000 ng of WT1+KTS, 300 ng of Nxf1 and 150 ng of Nxt1 or both depending on the experimental condition. At 72 hours post-transfection, supernatants were collected and analyzed for p24. The values shown are averages of duplicate transfections.

The MPMV-CTE containing reporter vector was transfected alone, with WT1+KTS, with Nxf1:Nxt1 and with both WT1 +KTS and Nxf1:Nxt1 in 293T cells. The RRE containing reporter vector was also used in the experiment as a negative control (Rev was not supplied in-*trans*). Cell supernatants were collected at 72 hours and a p24 ELISA assay was used to detect protein expression under the various conditions. Each transfection condition was performed in duplicate. Consistent with previously published results, the RRE alone (without Rev) did not generate detectable levels of p24 at 72 hours (< 1 ng/ml) (Figure 27 Panel B). Interestingly, when the MPMV-CTE constructs was transfected alone it also did not demonstrate detectable levels of p24, suggesting low levels of endogenous functional Nxf1/Nxt1 in the 293T cells. However, when the MPMV-CTE plasmid was transfected with WT1+KTS, detectable levels of p24 were achieved (26.7 ng/ml), though transfection with Nxf1/Nxt1 gave 10-fold higher levels (306 ng/ml).

However, when WT1 +KTS and Nxf1:Nxt1 were co-transfected with the GagPol-MPMV-CTE plasmid, the levels of p24 increased 10-fold again to 4,053 ng/ml. The data suggests that WT1 +KTS acts synergistically with Nxf1/Nxt1 to vastly increase protein production from the plasmid containing the CTE.

WT1 +KTS increases mRNA with retained introns in the cytoplasm

I next used the WT1+KTS RNA-seq data to discover if mRNA with retained introns were increased in the cytoplasm. As a control, I again used the empty vector RNA-seq data. Using the IRFinder pipeline [9] described in Chapter 2 of this thesis, I found that WT1+KTS significantly increased mRNA with retained introns in the cytoplasm as compared to the EV control. 1,194 introns were differentially expressed between conditions ($|\log2fc| > 1.5$, padj value < 0.05). Of these, 1,181 intron retention events from a total of 877 genes were increased with WT1+KTS expression (98.9%) (Appendix Table 10). Several introns demonstrated significant retention including an intron in the BBC3 gene which demonstrated a 270-fold change with WT1+KTS expression and an intron in the AT3 gene which demonstrated a 248-fold change with WT1+KTS expression. A GO enrichment analysis on the genes that demonstrated significant differential intron retention with WT1+KTS expression was performed (Figure 28, Panel A). There was once again a clear predominance on post-transcriptional regulatory pathways including mRNA binding and translation regulation. The genes with intron retention events found with WT1+KTS expression were also correlated with the direct RNA binding targets of WT1+KTS in the Hastie study [196]. Of the 877 genes that had at least one transcript with increased intron retention, 325 were reported to be direct RNA binding targets (Fisher's exact test, $p < 0.0001$). Additionally, of the 325 direct RNA binding targets of WT1 with increased intron retention, 23 had CTEs identified in the vector trap described in Chapter 2 of this thesis.

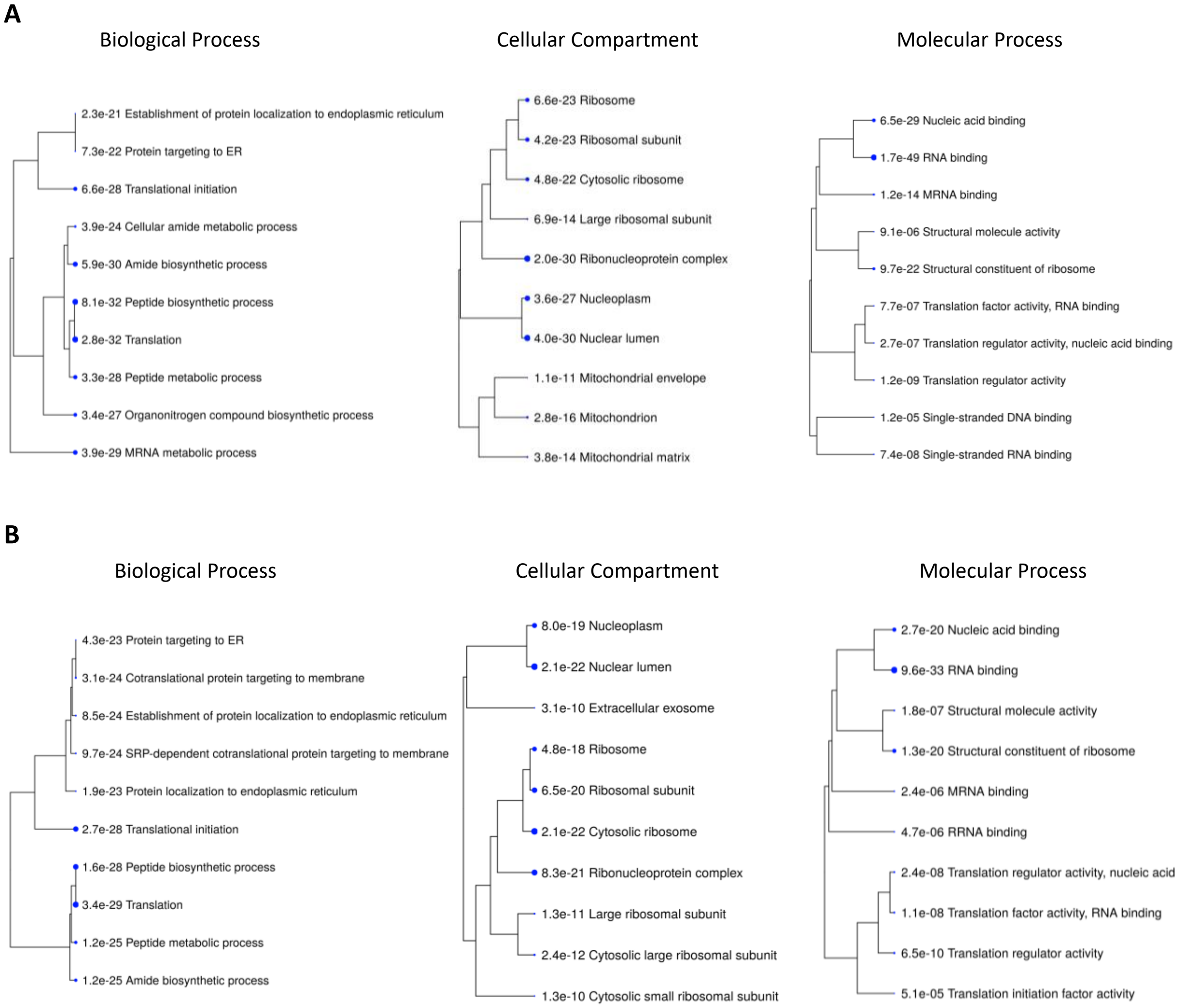
A
Biological Process
2.3e-21 Establishment of protein localization to endoplasmic reticulum
7.3e-22 Protein targeting to ER
6.6e-28 Translational initiation
3.9e-24 Cellular amide metabolic process
5.9e-30 Amide biosynthetic process
8.1e-32 Peptide biosynthetic process
2.8e-32 Translation
3.3e-28 Peptide metabolic process
3.4e-27 Organonitrogen compound biosynthetic process
3.9e-29 MRNA metabolic process
Cellular Compartment
6.6e-23 Ribosome
4.2e-23 Ribosomal subunit
4.8e-22 Cytosolic ribosome
6.9e-14 Large ribosomal subunit
2.0e-30 Ribonucleoprotein complex
3.6e-27 Nucleoplasm
4.0e-30 Nuclear lumen
1.1e-11 Mitochondrial envelope
2.8e-16 Mitochondrion
3.8e-14 Mitochondrial matrix
Molecular Process
6.5e-29 Nucleic acid binding
1.7e-49 RNA binding
1.2e-14 MRNA binding
9.1e-06 Structural molecule activity
9.7e-22 Structural constituent of ribosome
7.7e-07 Translation factor activity, RNA binding
2.7e-07 Translation regulator activity, nucleic acid binding
1.2e-09 Translation regulator activity
1.2e-05 Single-stranded DNA binding
7.4e-08 Single-stranded RNA binding
B
Biological Process
4.3e-23 Protein targeting to ER
3.1e-24 Cotranslational protein targeting to membrane
8.5e-24 Establishment of protein localization to endoplasmic reticulum
9.7e-24 SRP-dependent cotranslational protein targeting to membrane
1.9e-23 Protein localization to endoplasmic reticulum
2.7e-28 Translational initiation
1.6e-28 Peptide biosynthetic process
3.4e-29 Translation
1.2e-25 Peptide metabolic process
1.2e-25 Amide biosynthetic process
Cellular Compartment
8.0e-19 Nucleoplasm
2.1e-22 Nuclear lumen
3.1e-10 Extracellular exosome
4.8e-18 Ribosome
6.5e-20 Ribosomal subunit
2.1e-22 Cytosolic ribosome
8.3e-21 Ribonucleoprotein complex
1.3e-11 Large ribosomal subunit
2.4e-12 Cytosolic large ribosomal subunit
1.3e-10 Cytosolic small ribosomal subunit
Molecular Process
2.7e-20 Nucleic acid binding
9.6e-33 RNA binding
1.8e-07 Structural molecule activity
1.3e-20 Structural constituent of ribosome
2.4e-06 MRNA binding
4.7e-06 RRNA binding
2.4e-08 Translation regulator activity, nucleic acid
1.1e-08 Translation factor activity, RNA binding
6.5e-10 Translation regulator activity
5.1e-05 Translation initiation factor activity

Figure 28: WT1 +KTS plays a significant role in cytoplasmic intron retention. Genes with significant intron retention discovered in the IRFinder pipeline following comparison of Panel A- cytoplasmic WT1+KTS expression with cytoplasmic EV control; or Panel B- cytoplasmic WT1+KTS and Nxf1:Nxt1with cytoplasmic EV; were used for a Gene Ontology Enrichment Analysis performed using the ShinyGO visualization platform. Gene pathways that demonstrated an enrichment false discovery rate < 0.05 were included. Displayed above are the top 10 gene pathways (organized by FDR significance) in each respective category of the GO analysis- Biological Process, Cellular Component and Molecular Process.

WT1 +KTS does not affect mRNA with retained introns in total mRNA

As described above, total mRNA was also isolated from WT1+KTS and EV transfected 293T cells with the same transfection and sequencing parameters as the cytoplasmic data. I first compared total RNA-seq data from WT1+KTS transfected cells to cytoplasmic RNA-seq data from WT1+KTS transfected cells. Using the IRFinder pipeline, there were 3,355 differentially expressed introns, of which 3,331 introns were increased in total mRNA (99.3%). Similar to the Nxf1 RNA-seq data from Chapter 2, this data supports that a large number of mRNA with retained introns are maintained in the nucleus, likely a result of detention at the nuclear pore.

When total WT1+KTS was compared to total EV data using the IRFinder pipeline, there were only 142 differentially retained introns, 104 of which were increased with WT1+KTS expression (Appendix Table 11). In reference to the ~1,200 significantly retained introns when the cytoplasmic WT1+KTS data was analyzed, the total RNAseq data suggests that the role of WT1+KTS on intron containing mRNA is not likely to be a process that occurs in the nucleus. Though the mechanism cannot be confirmed by this data, the data suggests that WT1+KTS is not directly affecting alternative splicing of mRNA with retained introns. The difference between the total and cytoplasmic WT1+KTS RNA-seq data cannot distinguish between roles in nucleocytoplasmic export (as suspected with Nxf1) or in RNA stabilization as may be expected given previous publications showing WT1+KTS increases polyribosomal association of mRNA [27, 222, 231]. Further investigations, specifically analyzing polyribosomal RNA, is needed following WT1+KTS expression to look at direct effects on translation.

WT1 +KTS and co-transfected Nxf1/Nxt1 increases mRNA with retained introns in the cytoplasm

I next performed a separate transfection experiment were WT1+KTS and Nxf1:Nxt1 were co-transfected together in 293T cells. Cytoplasmic and total RNA was again isolated at 72 hours and sent to Novogene for RNA-sequencing using the same protocol used for the other experimental conditions in this chapter. The experiment generated approximately 40M stranded, pair-end reads per replicate (for both cytoplasmic and total). When the WT1+KTS and co-transfected Nxf1:Nxt1 cytoplasmic mRNA-seq data was compared to the cytoplasmic EV data with the IRFinder pipeline, there was once again many mRNA with retained introns in the cytoplasm, though interestingly not as dramatic as with WT1+KTS alone. There were 514 transcripts with retained introns from 385 genes (Appendix Table 12). Of the 514 transcripts with retained introns, 495 demonstrated increased cytoplasmic intron retention with WT1+KTS and Nxf1:Nxt1 (96.3%). When the genes with intron retention discovered in the WT1 +KTS and Nxf1:Nxt1 co-transfection experiment were used to perform a GO enrichment analysis (Figure 28, Panel B), there was again a predominance on post-transcriptional regulatory processes including RNA binding, ribosomal cellular compartment and translation initiation.

Interestingly, when total WT1+KTS and co-transfected Nxf1:Nxt1 RNAseq data was compared to total EV control data using the IRFinder pipeline, there were a total of 815 intron retention from 633 genes (Appendix Table 13). Of the 815 significantly retained intron 697 were increased with WT1+KTS and Nxf1:Nxt1 co-transfection (85%). This is a slightly different finding then when total RNA-seq data from Nxf1:Nxt1 and WT1+KTS was individually compared to total EV, which both noted very little intron retention in total RNA. Further investigations into a possible joint role of Nxf1 and WT1+KTS on RNA regulation within the nucleus including on splicing or nuclear trafficking, is necessary.

Discussion

The role of the WT1 gene in both health and disease has been heavily investigated over the last 30 years [59]. The gene was classified as one of the original tumor suppressor genes in Wilms' tumor in 1990, given the need for inactivating mutations in both alleles for tumors to develop [212, 213]. It is now understood to play a key role in cellular differentiation and development specifically regulating the mesenchyme to epithelial transitions (MET) and epithelial to mesenchymal transitions (EMT) [59]. WT1 is a complex gene with numerous mRNA isoforms with distinct functions. The WT1 gene has a four Cys2-His2 zinc finger binding domain at its carboxy terminus. One of the key functionally important isoforms is the inclusion or exclusion of 3 amino-acids (Lys-Thr-Ser) in between the 3rd and 4th zinc finger. The biological importance of these two WT1 isoforms was first discovered in the late 1990s. It was noted that splice variants in the WT1 gene of patients with Frasier syndrome caused a predominance of the -KTS isoform leading to the disease state which includes Wilms' tumor development and nephrosclerosis [216]. In early 2000s, Hammes and colleagues noted that mice expressing only -KTS or +KTS isoforms died neonatally through incomplete kidney development, suggesting both isoforms were independent and essential [232].

Specific NMR analysis has now demonstrated that the insertion of the 3 amino acids (KTS), increases the flexibility of the linker between the 3rd and 4th zinc finger, abrogating the DNA binding ability of the 2nd-4th zinc fingers [227]. Numerous subsequent investigations have highlighted the DNA binding capability of the -KTS isoform and its functional importance [225-227]. CHIP and CHIP-seq analyses have demonstrated several thousand WT1 transcriptional targets that regulate kidney development, almost all of which are specific to the -KTS isoform [218, 233-235]. Interestingly, WT1 has both a transcriptional activator and repressor domains and can function as either depending on its binding partners [236]. This has led to surprisingly diverse effects in numerous cancers, including several in which WT1 appears to act as an oncogene [237, 238]. The specific pattern of WT1 expression in cancer has led to be named one of the most important immunotherapeutic targets in cancer research by the National Institute of Health [239].

Specific to this thesis is the numerous observations that WT1+KTS isoform has a role in post-transcription regulation. An investigation in 1995 by Larsson and colleagues noted that the +KTS isoform specifically interacts and localizes with splicing factors in kidney cells [224]. It was further demonstrated that WT1+KTS binds splicing factors U2AF2 and is incorporated into functional spliceosomes in cell free systems [240]. This was followed by the observation that WT1 +KTS binds RNA with significantly higher affinity than -KTS isoform as a result of the atypical 1st zinc finger [241].

Later studies in the 2000s noted that WT1 undergoes CRM1 independent nucleocytoplasmic shuttling and specifically localizes to actively translating polyribosomes [222, 231]. Our previous data utilizing the WT1 +KTS isoform also suggests a very similar role for this isoform in the translation of mRNA with a retained intron [27]. We demonstrated that WT1+KTS interacts with mRNA containing a CTE and increases polyribosomal association of intron containing mRNA. Furthermore, in a previous publication, our data noted that with the addition of EDTA, the WT1+KTS isoform specifically localized to the small ribosomal protein, suggesting a dual role in mRNA and ribosomal association and binding. This led us to hypothesize that WT1+KTS promotes the translation of specific alternatively spliced transcripts. Further support of a direct RNA binding mechanism came following a UV crosslinking and sequencing experiment to identify mRNAs that interact with WT1 [196]. The experiment noted approximately 2,000 mRNA targets which specifically clustered in cell processes of cell adhesion and cell migration.

This investigation further showed that most of the WT1 binding occurred at the 3'UTR and possibly regulated RNA stability [196].

The data from our current investigation is consistent with the preliminary investigations of the post-transcriptional role of WT1+KTS. The fact that 293T cells do not express WT1 allowed us to isolate the transcriptional effects of the WT1+KTS expression in this model cell system. Our differential gene expression analysis following WT1+KTS expression was less than 5 genes, suggesting that the isoform does not have a major role as a transcriptional regulator as does the WT1-KTS isoform, which has been previously suggested [218, 233-235]. Furthermore, our differential transcript expression noted numerous transcripts which were increased in the cytoplasm following WT1+KTS expression. Very similar to the RNA binding targets of WT1 noted in the Hastie immunoprecipitation experiment, WT1+KTS expression affected transcripts with roles in post-transcriptional regulation including RNA binding, cellular localization and ribonucleotide binding [27, 59].

As the WT1+KTS isoform appears to regulate transcripts that are involved in post-transcriptional processes, we were interested in discovering if a portion of the cellular effects seen by WT1+KTS were a result of direct RNA binding targets or through secondary post-transcriptional effects. When we compared our DTE list with the direct binding targets from the immunoprecipitation experiment performed by the Hastie lab, we noted that a quarter of our differentially expressed transcripts were also direct binding targets. When we reviewed the gene enrichment of this overlap, we once again found that WT1+KTS targets are involved in RNA splicing, nucleocytoplasmic export and protein binding. Additionally, we noted over 1,300 genes that had differential transcript isoform ratios in the cytoplasm as a result of WT1+KTS expression.

The data on WT1+KTS mediating an isoform switch of c-Myc is itself very intriguing and highlights the importance of post-transcriptional role of the WT1+KTS isoform. c-Myc is a well describe transcription factor with essential roles in cell proliferation and development [242]. Together with its dimerization partner MAX, a basic helix loop helix leucine zipper protein, c-Myc is a master regulator affecting the expression patterns of thousands of genes [243-246]. It is also a well described oncogene and dysregulation of the gene causes aggressive disease [242]. c-Myc has approximately 8 different mRNA transcripts described in Ensembl and 3 major protein isoforms termed Myc1, Myc2 and a substantially shorter protein called MycS, which interestingly appear to be regulated by alternate translation start sites, differing only at the N-terminal end of the protein [247]. The different proteins have significantly different cellular functions including c-Myc2 which appears predominately to mediate transcriptional activation, while c-Myc1 *arrests* cell growth and leads to apoptosis [229, 230]. In this investigation, we demonstrated that WT1+KTS expression caused an isoform switch in the cytoplasm from c-Myc2 to c-Myc1.

The isoform switch of c-Myc by WT1+KTS expression is also intriguing as previous investigations have demonstrated that WT1 may be a transcriptional regulator of c-Myc [248]. Specifically, WT1 was shown to bind the second major transcriptional start site of c-Myc leading to upregulation of c-Myc in multiple breast cancer cell lines [249]. WT1 was also shown to bind and promote c-Myc expression in K-Ras mutant non-small cell lung cancer [250]. As the majority of the previous investigations did not distinguish between WT1 isoforms nor c-Myc isoforms, some of the effect of the WT1- c-Myc axis that has been investigated may result in part due to the post-transcriptional regulation of c-Myc by the WT1+KTS isoform.

In addition to the differential transcript expression results, the cytoplasmic expression of mRNA isoforms with retained introns was also strong following WT1+KTS transfection. There were 1,300 intron retention events, 98% of which were increased in the cytoplasmic with WT1+KTS expression. In comparison, Nxf1:Nxt1 over-expression experiment performed in chapter 2 of this thesis, increased the retention of 400 introns and the co-transfection of WT1(+KTS) and Nxf1:Nxt1 increased 500 mRNA isoforms with retained introns. The findings of WT1+KTS increasing cytoplasmic intron retention of mammalian genes is not an unexpected finding given the previously published data demonstrating the strong function of WT1+KTS on the MPMV-CTE [27]. However, our previously published data, which was re-confirmed by the p24 results in 293T cells from this chapter, suggested that WT1(+KTS) likely worked at the level of translation, making the cytoplasmic mRNA data interesting. Whether the cytoplasmic results were reflective of increased mRNA stabilization or conversely were associated with nuclear export cannot be stated from the data presented in these experiments. However, it does support the potential role of mRNA with retained introns involved in regulating post-transcriptional processes, which was also supported by our Nxf1 over-expression data from Chapter 2. It additionally supports further investigations into the translation of mRNA with retained introns in the presence of WT1+KTS expression.

Conclusion

Our data demonstrate that WT1+KTS appears to increase a specific subset of differentially expressed transcripts which regulate post-transcriptional processes including RNA-processing. Furthermore, WT1+KTS appears to differentially affect mRNA isoform ratios in the cytoplasm including an isoform switch in c-MYC, though this requires experimental validation. Furthermore, WT1+KTS appears to increase, through an unknown mechanism, mRNA isoform with retained introns in the cytoplasm. The genes with retained introns appear to cluster in genes involved in post-transcriptional regulatory processes.

Methods

Plasmids

All plasmids utilized in this chapter follow the nomenclature pHRXXXX for identification and eases of requests. The major plasmids utilized include WT1(+exon 5, +KTS) pHR3056, which was a gift from Mike Ladomery and Nick Hastie and have been previously described [251]. The addition plasmids used include pCMV-Empty Vector (pHR0016), pCMV-NXF1 (pHR3704) and pCMV-NXT1 (pHR2415) and have also been previously described [209]. The reporter vector utilized in the p24 analysis is pCMV-GagPol-MPMV-CTE (pHR1361) and pCMV-GagPol-RRE (pHR3442).

Cell Lines and Cell Transfections

293T cells were maintained in DMEM, 10% Bovine Calf Serum (BCS) and 50 mg/ml gentamicin. Transient transfections were performed using a lipofectamine3000 protocol as suggest by the manufacturer's instructions (Thermofisher Scientific). Specifically, 293T cells (10 cm plates, 3x10^6 cells, 10 ml media) were transfected with 15 ug of WT1(+KTS), or 15 ug of WT1(+KTS) and 5 ug of Nxf1 and 10 ug of Nxt1, or 15 ug of pCMV-Empty vector. Cells were harvested 72 hours post-transfection. Each experimental condition was performed in triplicate in two separate experiments (6 total transfected plates per experimental condition). 3 cell plates per transfection condition were utilized for either cytoplasmic or total RNA isolation.

A phenol: chloroform extraction protocol was utilized for both total and cytoplasmic RNA isolation as previously described [209]. Key steps to the cytoplasmic protocol include washing pelleted cells with ice-cold phosphate-buffered saline (PBS) followed by resuspension in a reticulocyte standard buffer (RSB) (10 mM Tris-HCl (pH 7.4), 10 mM NaCl, 1.5 mM MgCl2) followed by the addition of RSB with 0.1% IgePal in equal volume to perform cell lysis. Cell nuclei were pelleted out using two sequential centrifugations at 13,000 rpm in an Eppendorf centrifuge at +4C. 2 PK buffer (200 mM Tris-HCl (pH 7.5), 25 mM EDTA, 300 mM NaCl, 2% SDS, 400 ug of proteinase K/ml) was added to cell lysates (with nuclei cleared) and incubated at 37C for 30 min. Cytoplasmic RNA was then extracted from the lysate using a phenol and chloroform-isoamyl alcohol (24:1) in a 1:1 volume followed by 2 separate extractions with chloroform-isoamyl alcohol (24:1). RNA was precipitated using 200 proof ethanol (-80C) at a 2.5 volume with the addition of sodium acetate (3M, pH 5.5- final concentration 0.3 M. RNA precipitated in EtOH was stored at -80C. Key steps for the total RNA extraction including washing pelleted cells twice with ice cold PBS. Cells were lysed with lysis buffer containing 0.2M Tris-HCL (pH 7.5), 0.2 M NaCl, 1.5 mM $MgCl_2$, 2% sodium docdecyl sulfate (SDS) and 200 ug of proteinase K per ml and incubated for 1 hour at 45C.

RNA-EtOH slurry was pelleted for 1 hour at 4C at 3,000 RPM (3014g) in a Baxter Cryofuge 6000 centrifuge. RNA pellet was washed twice with 75% ethanol and air dried twice. Total RNA additionally DNase digested (RQ1 RNase-free DNase, Promega) then extracted as above. RNA was resuspended in 10 ul of RNase/DNase free water. RNA concentrations were determined by Qubit RNA HS assay kit (Qubit 3.0 fluorometer, Thermofisher). RNA quality was determined by RNA Tape station. PolyA+ RNA was isolated from the cytoplasmic RNA using poly-T magnetic bead-based protocol according to the manufactures protocol (NEXTFLEX Poly(A) Beads 2.0 Kit, PerkinElmer Inc).

Isolated RNA was subjected to both short read sequencing. Short read sequencing was performed at Novogene (Beijing, China). Isolated polyA+ RNA was shipped to Novogene (on dry ice) who reconfirmed RNA quality and prepared stranded cDNA libraries. Libraries were sequenced on an Illumina HiSeq3000 system to generate paired-end 150 bp reads to a sequencing depth of approximately 40M reads per sample.

P24 ELISA Experiment

The P24 ELISA assay utilized in this chapter has been previously described [209]. For the specific experiment, 293T cells (5.5x10^5 cells in 6-well plate, 2 ml media/well) were transfected with 1,000 ng of pCMV-GagPol reporter plasmid with either the RRE or the MPMV-CTE in addition to 1,000 ng of WT1+KTS, 300 ng of Nxf1 and 150 ng of Nxt1 or both depending on the experimental condition. At 72 hours post-transfection, supernatants were collected and analyzed for p24. Experiments were performed in duplicate.

Bioinformatic Analysis for short read RNA-seq data

Sequencing reads were trimmed and quality controlled utilizing the programs Trimmomatic and FastQC (http://www.bioinformatics.bbsrc.ac.uk/projects/fastqc) [146]. We used the splice aware differential gene expression pipeline- HiSAT2, Stringtie and Ballgown with recommendation conditions (genome and annotation file used were Ensembl Human Genome Hg38.v37 and corresponding GTF file) [154]. We additionally performed a differential transcript expression and differential transcript usage analysis using Salmon and DESeq2 and DRIMSeq according to the suggested parameters (Ensembl GRCh38.95 cDNA fasta) [210]. Lastly, we utilized IRFinder for intron retention discovery and differential analysis according to specifications [9]. IRFinder is based on the STAR alignment algorithm (version STAR-2.7.3a). RNA-seq reads were aligned to Ensembl Human Genome Hg38.v37 and corresponding GTF file. IRFinder calculates an IRatio across different introns. Differential expression of individual IRatio between conditions was performed using the IRFinder provided DESeq2 R script under recommend specifications. Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) over-representation functional analysis was performed on Genes that contained intron retention events. Gene pathways were considered enriched if they had a p-adj value < 0.05. GO and KEGG analysis was performed using the ShinyGO package for visualization and the clusterProfiler package in R [151, 174].

VI. Conclusion

In this thesis, I provide additional evidence that the post-transcriptional regulation of alternatively spliced transcripts, specifically mRNA with retained introns, have important biological effects on mammalian systems. The data presented in this thesis is the first evidence that Human Endogenous Retrovirus-K expression is upregulated in fetal solid organ tumors. Furthermore, the specificity of HERV-K expression in tumor tissue combined with the very minimal expression in normal tumor controls, makes HERV-K an intriguing target for immunotherapy and tumor markers in both hepatoblastoma and Wilms' tumor. Future studies regarding the effects of Rec expression in Wilms' tumor may prove very exciting. The data from this thesis also demonstrates that CTEs are a conserved molecular mechanism by which intron containing mRNA can gain nuclear export competence in mammalian systems. The data also demonstrate that Nxf1 selectively increases cytoplasmic intron retention in numerous mammalian genes and has a close overlap with genes that contain CTEs. Lasty, I show data that demonstrates the WT1+KTS isoform increases differentially expressed transcripts in the cytoplasm which appear to regulate post-transcriptional processes. Furthermore, the data in this thesis show that WT1+KTS has a clear role in cytoplasmic intron retention.

VII. Future Work

There are important areas of on-going research in our lab regarding the investigations presented in each chapter of this thesis. Regarding the work on the HERV-K annotation, an immediate goal is to experimentally confirm which proviruses are capable of producing individual viral proteins. Our lab has also already begun experimentally testing our predicted HERV-K Rec transcripts to discover which are cable of producing functional export proteins. From a bioinformatic perspective, the utility of the full HERV-K annotation will need to be further established by using long read RNA-seq data in a controlled experiment. We are currently considering a cancer vs. non-cancer cell-line experiment to demonstrate the utility of the HERV-K annotation using long reads. The increased HERV-K expression seen in fetal tumors is very intriguing, but follow up investigations into protein expression is necessary. Additionally, direct correlation of HERV-K expression profiles with tumor subtypes may prove very informative, especially utilizing the Wilms' tumor dataset.

Regarding the CTE project, the immediate next step in our lab is to investigate the RNA secondary and tertiary structures of the identified functional cCTEs to determine if structural homology exists between the functional elements, which would suggest a structurally specific domain. As noted in our methods the identified CTEs are currently part of cDNA fragments that range from 500-1,000 bp. From our understanding of viral CTEs and the Nxf1-CTE, the CTE itself is likely approximately 300 bp long. One initial option will be to trim the existing functional CTEs and identify the 300 bp sequence that exists as a functional CTE, prior to secondary RNA analysis. Additionally, it will be necessary to perform polyribosomal and Ribo-Seq analysis following Nxf1 over-expression to see if the IR targets identified in the cytoplasm are indeed translated.

Lastly, there is significant additional work to perform regarding the WT1+KTS investigation. The data from this thesis supports but does not confirm previous work that WT1+KTS traffics a subset of mRNAs, including mRNAs with retained introns, from the spliceosome to the ribosome. Validating this trafficking pattern through immunofluorescence microscopy is an important next experiment. Additionally, we have data that suggest WT1+KTS protein may directly bind to ribosomal subunits- promoting translation of specific mRNAs. Confirming this protein- protein interaction through a co-immunoprecipitation or pull-down assay would be extremely informative. Similar to the Nxf1 experiment, examining polyribosomal RNA following WT1+KTS expression is immediately necessary to show that the differential transcript expression we found in cytoplasmic RNA is also demonstrated at the protein level.

VIII. Acknowledgements

I would like to thank my primary research mentors Marie-Louise Hammarskjold and David Rekosh for sharing their infectious love of science and discovery. They taught me to be a perpetually curious scientist and to value quality over quantity. I would like to thank Craig Slingluff for accepting me on his surgical oncology training grant, truly at the last possible moment, as well as his perspective balancing the rigorous of science, surgery and sailing. I would additionally like to thank my committee members, Sara Rasmussen, Aakrosh Ratan, Stefan Bekiranov and the chair of my committee, John Bushweller, for the enormous amount of time and energy they dedicated to my academic progress. Lastly, I would like to thank Avril Somlyo and the Department of Molecular Physiology and Biological Physics for championing the clinician-scientist program at the University of Virginia.

IX. References

1. Wang, E.T., et al., *Alternative isoform regulation in human tissue transcriptomes.* Nature, 2008. **456**(7221): p. 470-6.
2. Berget, S.M., C. Moore, and P.A. Sharp, *Spliced segments at the 5' terminus of adenovirus 2 late mRNA. 1977.* Rev Med Virol, 2000. **10**(6): p. 356-62; discussion 355-6.
3. Nilsen, T.W. and B.R. Graveley, *Expansion of the eukaryotic proteome by alternative splicing.* Nature, 2010. **463**(7280): p. 457-63.
4. Baralle, F.E. and J. Giudice, *Alternative splicing as a regulator of development and tissue identity.* Nature Reviews Molecular Cell Biology, 2017. **18**(7): p. 437-451.
5. Rekosh, D. and M.-L. Hammarskjold, *Intron retention in viruses and cellular genes: Detention, border controls and passports.* Wiley interdisciplinary reviews. RNA, 2018. **9**(3): p. e1470.
6. Ner-Gaon, H., et al., *Intron retention is a major phenomenon in alternative splicing in Arabidopsis.* Plant J, 2004. **39**(6): p. 877-85.
7. Hammarskjöld, M.L., *Regulation of retroviral RNA export.* Seminars in Cell & Developmental Biology, 1997. **8**(1): p. 83-90.
8. Braunschweig, U., et al., *Widespread intron retention in mammals functionally tunes transcriptomes.* Genome Research, 2014. **24**(11): p. 1774-1786.
9. Middleton, R., et al., *IRFinder: assessing the impact of intron retention on mammalian gene expression.* Genome Biology, 2017. **18**(1): p. 51.
10. Edwards, C.R., et al., *A dynamic intron retention program in the mammalian megakaryocyte and erythrocyte lineages.* Blood, 2016.

11. Memon, D., et al., *Hypoxia-driven splicing into noncoding isoforms regulates the DNA damage response.* NPJ Genom Med, 2016. **1**: p. 16020.

12. Pimentel, H., et al., *A dynamic intron retention program enriched in RNA processing genes regulates gene expression during terminal erythropoiesis.* Nucleic Acids Research, 2016. **44**(2): p. 838-851.

13. Ullrich, S. and R. Guigo, *Dynamic changes in intron retention are tightly associated with regulation of splicing factors and proliferative activity during B-cell development.* Nucleic Acids Res, 2020. **48**(3): p. 1327-1340.

14. Yap, K., et al., *Coordinated regulation of neuronal mRNA steady-state levels through developmentally controlled intron retention.* Genes & Development, 2012. **26**(11): p. 1209-1223.

15. Buckley, P.T., et al., *Cytoplasmic intron sequence-retaining transcripts can be dendritically targeted via ID element retrotransposons.* Neuron, 2011. **69**(5): p. 877-884.

16. Mauger, O., F. Lemoine, and P. Scheiffele, *Targeted Intron Retention and Excision for Rapid Gene Regulation in Response to Neuronal Activity.* Neuron, 2016. **92**(6): p. 1266-1278.

17. Adusumalli, S., et al., *Increased intron retention is a post-transcriptional signature associated with progressive aging and Alzheimer's disease.* Aging Cell, 2019. **18**(3): p. e12928.

18. Dvinge, H. and R.K. Bradley, *Widespread intron retention diversifies most cancer transcriptomes.* Genome Med, 2015. **7**(1): p. 45.

19. Casadei, S., et al., *Characterization of splice-altering mutations in inherited predisposition to cancer.* Proc Natl Acad Sci U S A, 2019.

20. Bray, M., et al., *A small element from the Mason-Pfizer monkey virus genome makes human immunodeficiency virus type 1 expression and replication Rev-independent.* Proc Natl Acad Sci U S A, 1994. **91**(4): p. 1256-60.

21. Hammarskjold, M.L., et al., *Regulation of human immunodeficiency virus env expression by the rev gene product.* J Virol, 1989. **63**(5): p. 1959-66.

22. Hammarskjöld, M.L., et al., *Human immunodeficiency virus env expression becomes Rev-independent if the env region is not defined as an intron.* J Virol, 1994. **68**(2): p. 951-8.

23. Löwer, R., et al., *Identification of a Rev-related protein by analysis of spliced transcripts of the human endogenous retroviruses HTDV/HERV-K.* J Virol, 1995. **69**(1): p. 141-9.

24. Magin-Lachmann, C., et al., *Rec (formerly Corf) function requires interaction with a complex, folded RNA structure within its responsive element rather than binding to a discrete specific binding site.* J Virol, 2001. **75**(21): p. 10359-71.

25. Li, Y., et al., *An NXF1 mRNA with a retained intron is expressed in hippocampal and neocortical neurons and is translated into a protein that functions as an Nxf1 cofactor.* Molecular Biology of the Cell, 2016. **27**(24): p. 3903-3912.

26. Coyle, J.H., et al., *Sam68 enhances the cytoplasmic utilization of intron-containing RNA and is functionally regulated by the nuclear kinase Sik/BRK.* Mol Cell Biol, 2003. **23**(1): p. 92-103.

27. Bor, Y.C., et al., *The Wilms' tumor 1 (WT1) gene (+KTS isoform) functions with a CTE to enhance translation from an unspliced RNA with a retained intron.* Genes Dev, 2006. **20**(12): p. 1597-608.

28. Chang, D.D. and P.A. Sharp, *Regulation by HIV Rev depends upon recognition of splice sites.* Cell, 1989. **59**(5): p. 789-795.

29. Legrain, P. and M. Rosbash, *Some cis- and trans-acting mutants for splicing target pre-mRNA to the cytoplasm.* Cell, 1989. **57**(4): p. 573-583.

30. Boutz, P.L., A. Bhutkar, and P.A. Sharp, *Detained introns are a novel, widespread class of post-transcriptionally spliced introns.* Genes & Development, 2015. **29**(1): p. 63-80.

31. Green, R.E., et al., *Widespread predicted nonsense-mediated mRNA decay of alternatively-spliced transcripts of human normal and disease genes.* Bioinformatics, 2003. **19 Suppl 1**: p. i118-21.

32. Kurosaki, T. and L.E. Maquat, *Nonsense-mediated mRNA decay in humans at a glance.* J Cell Sci, 2016. **129**(3): p. 461-7.

33. Hammarskjold, M.L., et al., *Northern Blot analysis of mRNA from mammalian polyribosomes.* Nature Methods, 2006.

34. Schmitz, U., et al., *Intron retention enhances gene regulatory complexity in vertebrates.* Genome Biol, 2017. **18**(1): p. 216.

35. Wong, J.J.L., et al., *Orchestrated intron retention regulates normal granulocyte differentiation.* Cell, 2013. **154**(3): p. 583-595.

36. Pimentel, H., J.G. Conboy, and L. Pachter, *Keep Me Around: Intron Retention Detection and Analysis.* arXiv:1510.00696 [q-bio], 2015.

37. Broseus, L. and W. Ritchie, *Challenges in detecting and quantifying intron retention from next generation sequencing data.* Computational and Structural Biotechnology Journal, 2020. **18**: p. 501-508.

38. Grabski, D.F., et al., *Intron retention and its impact on gene expression and protein diversity: A review and a practical guide.* Wiley Interdiscip Rev RNA, 2021. **12**(1): p. e1631.

39. Sultan, M., et al., *Influence of RNA extraction methods and library selection schemes on RNA-seq data.* BMC Genomics, 2014. **15**: p. 675.

40. Zhao, S., et al., *Evaluation of two main RNA-seq approaches for gene quantification in clinical RNA sequencing: polyA+ selection versus rRNA depletion.* Sci Rep, 2018. **8**(1): p. 4781.

41. Bentley, D.L., *Coupling mRNA processing with transcription in time and space.* Nat Rev Genet, 2014. **15**(3): p. 163-75.

42. Merkhofer, E.C., P. Hu, and T.L. Johnson, *Introduction to cotranscriptional RNA splicing.* Methods Mol Biol, 2014. **1126**: p. 83-96.

43. Zhao, S., et al., *Comparison of stranded and non-stranded RNA-seq transcriptome profiling and investigation of gene overlap.* BMC Genomics, 2015. **16**(1): p. 675.

44. Barman, P., D. Reddy, and S.R. Bhaumik, *Mechanisms of Antisense Transcription Initiation with Implications in Gene Expression, Genomic Integrity and Disease Pathogenesis.* Noncoding RNA, 2019. **5**(1).

45. Vanichkina, D.P., et al., *Challenges in defining the role of intron retention in normal biology and disease.* Seminars in Cell & Developmental Biology, 2018. **75**: p. 40-49.

46. Amarasinghe, S.L., et al., *Opportunities and challenges in long-read sequencing data analysis.* Genome Biology, 2020. **21**(1): p. 30.

47. Ranganathan, S., D. Lopez-Terrada, and R. Alaggio, *Hepatoblastoma and Pediatric Hepatocellular Carcinoma: An Update.* Pediatr Dev Pathol, 2019: p. 1093526619875228.

48. Kremer, N., A.E. Walther, and G.M. Tiao, *Management of hepatoblastoma: an update.* Curr Opin Pediatr, 2014. **26**(3): p. 362-9.

49. Mavila, N. and J. Thundimadathil, *The Emerging Roles of Cancer Stem Cells and Wnt/Beta-Catenin Signaling in Hepatoblastoma.* Cancers (Basel), 2019. **11**(10).

50. Wu, J.F., et al., *Prognostic roles of pathology markers immunoexpression and clinical parameters in Hepatoblastoma.* J Biomed Sci, 2017. **24**(1): p. 62.

51. Ruck, P., et al., *Hepatic stem-like cells in hepatoblastoma: expression of cytokeratin 7, albumin and oval cell associated antigens detected by OV-1 and OV-6.* Histopathology, 1997. **31**(4): p. 324-9.

52. Czauderna, P., et al., *Hepatoblastoma state of the art: pathology, genetics, risk stratification, and chemotherapy.* Curr Opin Pediatr, 2014. **26**(1): p. 19-28.

53. Meyers, R.L., et al., *Hepatoblastoma state of the art: pre-treatment extent of disease, surgical resection guidelines and the role of liver transplantation.* Curr Opin Pediatr, 2014. **26**(1): p. 29-36.

54. Feng, J., et al., *Assessment of Survival of Pediatric Patients With Hepatoblastoma Who Received Chemotherapy Following Liver Transplant or Liver Resection.* JAMA Netw Open, 2019. **2**(10): p. e1912676.

55. Carceller, A., et al., *Surgical resection and chemotherapy improve survival rate for patients with hepatoblastoma.* J Pediatr Surg, 2001. **36**(5): p. 755-9.

56. Treger, T.D., et al., *The genetic changes of Wilms tumour.* Nat Rev Nephrol, 2019.

57. Dome, J.S., et al., *Advances in Wilms Tumor Treatment and Biology: Progress Through International Collaboration.* J Clin Oncol, 2015. **33**(27): p. 2999-3007.

58. Phelps, H.M., et al., *Biological Drivers of Wilms Tumor Prognosis and Treatment.* Children (Basel), 2018. **5**(11).

59. Hastie, N.D., *Wilms' tumour 1 (WT1) in development, homeostasis and disease.* Development, 2017. **144**(16): p. 2862-2872.

60. Ilmer, M., et al., *Targeting the Neurokinin-1 Receptor Compromises Canonical Wnt Signaling in Hepatoblastoma.* Mol Cancer Ther, 2015. **14**(12): p. 2712-21.

61. Indersie, E., et al., *MicroRNA therapy inhibits hepatoblastoma growth in vivo by targeting beta-catenin and Wnt signaling.* Hepatol Commun, 2017. **1**(2): p. 168-183.

62. Lee, H., et al., *General paucity of genomic alteration and low tumor mutation burden in refractory and metastatic hepatoblastoma: comprehensive genomic profiling study.* Hum Pathol, 2017. **70**: p. 84-91.

63. Griffiths, D.J., *Endogenous retroviruses in the human genome sequence.* Genome Biol, 2001. **2**(6): p. Reviews1017.

64. Mayer, J., J. Blomberg, and R.L. Seal, *A revised nomenclature for transcribed human endogenous retroviral loci.* Mob DNA, 2011. **2**(1): p. 7.

65. Shin, W., et al., *Human-specific HERV-K insertion causes genomic variations in the human genome.* PLoS One, 2013. **8**(4): p. e60605.

66. Vargiu, L., et al., *Classification and characterization of human endogenous retroviruses; mosaic forms are common.* Retrovirology, 2016. **13**: p. 7.

67. Stoye, J.P., *Studies of endogenous retroviruses reveal a continuing evolutionary saga.* Nat Rev Microbiol, 2012. **10**(6): p. 395-406.

68. Schmitt, K., et al., *Transcriptional profiling of human endogenous retrovirus group HERV-K(HML-2) loci in melanoma.* Genome biology and evolution, 2013. **5**(2): p. 307-328.

69. Menendez, L., B.B. Benigno, and J.F. McDonald, *L1 and HERV-W retrotransposons are hypomethylated in human ovarian carcinomas.* Molecular cancer, 2004. **3**: p. 12.

70. Jern, P. and J.M. Coffin, *Effects of retroviruses on host genome function.* Annu Rev Genet, 2008. **42**: p. 709-32.

71. Buzdin, A., et al., *At least 50% of human-specific HERV-K (HML-2) long terminal repeats serve in vivo as active promoters for host nonrepetitive DNA transcription.* J Virol, 2006. **80**(21): p. 10752-62.

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

73. Samuelson, L.C., et al., *Retroviral and pseudogene insertion sites reveal the lineage of human salivary and pancreatic amylase genes from a single gene during primate evolution.* Molecular and cellular biology, 1990. **10**(6): p. 2513-2520.

74. Voisset, C., R.A. Weiss, and D.J. Griffiths, *Human RNA "rumor" viruses: the search for novel human retroviruses in chronic disease.* Microbiology and molecular biology reviews : MMBR, 2008. **72**(1): p. 157-96, table of contents.

75. Kury, P., et al., *Human Endogenous Retroviruses in Neurological Diseases.* Trends Mol Med, 2018. **24**(4): p. 379-394.

76. Downey, R.F., et al., *Human endogenous retrovirus K and cancer: Innocent bystander or tumorigenic accomplice?* International journal of cancer.Journal international du cancer, 2014.

77. Nakagawa, K. and L.C. Harrison, *The potential roles of endogenous retroviruses in autoimmunity.* Immunological reviews, 1996. **152**: p. 193-236.

78. Bergallo, M., et al., *CMV induces HERV-K and HERV-W expression in kidney transplant recipients.* Journal of Clinical Virology, 2015. **68**((Mareschi K., katia.mereschi@unito.it; Fagioli F., franca.fagioli@unito.it) Pediatric Onco-Hematology, Stem Cell Transplantation and Cellular Therapy Division, City of Science and Health of Turin, Regina Margherita Children's Hospital, Turin, Italy): p. 28-31.

79. van der Kuyl, A.C., *HIV infection and HERV expression: a review.* Retrovirology, 2012. **9**: p. 6.

80. Perron, H., et al., *Human endogenous retrovirus type W envelope expression in blood and brain cells provides new insights into multiple sclerosis disease.* Mult Scler, 2012. **18**(12): p. 1721-36.

81. Rolland, A., et al., *The envelope protein of a human endogenous retrovirus-W family activates innate immunity through CD14/TLR4 and promotes Th1-like responses.* J Immunol, 2006. **176**(12): p. 7636-44.

82. Curtin, F., et al., *A placebo randomized controlled study to test the efficacy and safety of GNbAC1, a monoclonal antibody for the treatment of multiple sclerosis - Rationale and design.* Mult Scler Relat Disord, 2016. **9**: p. 95-100.

83. Humer, J., et al., *Identification of a melanoma marker derived from melanoma-associated endogenous retroviruses.* Cancer research, 2006. **66**(3): p. 1658-1663.

84. Schiavetti, F., et al., *A human endogenous retroviral sequence encoding an antigen recognized on melanoma by cytolytic T lymphocytes.* Cancer research, 2002. **62**(19): p. 5510-6.

85. Krishnamurthy, J., et al., *Genetic engineering of T cells to target HERV-K, an ancient retrovirus on melanoma.* Clinical Cancer Research, 2015. **21**(14): p. 3241-3251.

86. Wang-Johanning, F., et al., *Immunotherapeutic potential of anti-human endogenous retrovirus-K envelope protein antibodies in targeting breast tumors.* Journal of the National Cancer Institute, 2012. **104**(3): p. 189-210.

87. Zhou, F., et al., *Chimeric antigen receptor T cells targeting HERV-K inhibit breast cancer and its metastasis through downregulation of Ras.* Oncoimmunology, 2015. **4**(11): p. e1047582-e1047582.

88. Meyer, T.J., et al., *Endogenous Retroviruses: With Us and against Us.* Front Chem, 2017. **5**: p. 23.

89. Grabski, D.F., et al., *Close to the Bedside: A Systematic Review of Endogenous Retroviruses and Their Impact in Oncology.* J Surg Res, 2019. **240**: p. 145-155.

90. Blomberg, J., et al., *Classification and nomenclature of endogenous retroviral sequences (ERVs): problems and recommendations.* Gene, 2009. **448**(2): p. 115-23.

91. Hohn, O., K. Hanke, and N. Bannert, *HERV-K(HML-2), the Best Preserved Family of HERVs: Endogenization, Expression, and Implications in Health and Disease.* Front Oncol, 2013. **3**: p. 246.

92. Sverdlov, E.D., *Retroviruses and primate evolution.* Bioessays, 2000. **22**(2): p. 161-71.

93. Subramanian, R.P., et al., *Identification, characterization, and comparative genomic distribution of the HERV-K (HML-2) group of human endogenous retroviruses.* Retrovirology, 2011. **8**: p. 90.

94. Denne, M., et al., *Physical and functional interactions of human endogenous retrovirus proteins Np9 and rec with the promyelocytic leukemia zinc finger protein.* J Virol, 2007. **81**(11): p. 5607-16.

95. Chen, T., et al., *The viral oncogene Np9 acts as a critical molecular switch for co-activating β-catenin, ERK, Akt and Notch1 and promoting the growth of human leukemia stem/progenitor cells.* Leukemia, 2013. **27**(7): p. 1469-78.

96. Grow, E.J., et al., *Intrinsic retroviral reactivation in human preimplantation embryos and pluripotent cells.* Nature, 2015. **522**(7555): p. 221-5.

97. Muster, T., et al., *An endogenous retrovirus derived from human melanoma cells.* Cancer Res, 2003. **63**(24): p. 8735-41.

98. Wildschutte, J.H., et al., *Discovery of unfixed endogenous retrovirus insertions in diverse human populations.* Proc Natl Acad Sci U S A, 2016. **113**(16): p. E2326-34.

99. Pisano, M.P., N. Grandi, and E. Tramontano, *High-Throughput Sequencing is a Crucial Tool to Investigate the Contribution of Human Endogenous Retroviruses (HERVs) to Human Biology and Development.* Viruses, 2020. **12**(6).

100. Fuentes, D.R., T. Swigut, and J. Wysocka, *Systematic perturbation of retroviral LTRs reveals widespread long-range effects on human gene regulation.* Elife, 2018. **7**.

101. Montesion, M., et al., *Promoter expression of HERV-K (HML-2) provirus-derived sequences is related to LTR sequence variation and polymorphic transcription factor binding sites.* Retrovirology, 2018. **15**(1): p. 57.

102. Gray, L.R., et al., *HIV-1 Rev interacts with HERV-K RcREs present in the human genome and promotes export of unspliced HERV-K proviral RNA.* Retrovirology, 2019. **16**(1): p. 40.

103. Muster, T., et al., *An endogenous retrovirus derived from human melanoma cells.* Cancer Res, 2003. **63**(24): p. 8735-41.

104. Zhou, F., et al., *Activation of HERV-K Env protein is essential for tumorigenesis and metastasis of breast cancer cells.* Oncotarget, 2016. **7**(51): p. 84093-84117.

105. Ma, W., et al., *Human Endogenous retroviruses-k (HML-2) expression is correlated with prognosis and progress of hepatocellular carcinoma.* BioMed Research International, 2016. **2016**((Ding L.; Zhou F.) Department of Clinical Hematology, Zhongnan Hospital, Wuhan University, Wuhan, China).

106. Chiappinelli, K.B., et al., *Inhibiting DNA Methylation Causes an Interferon Response in Cancer via dsRNA Including Endogenous Retroviruses.* Cell, 2015. **162**(5): p. 974-86.

107. Roulois, D., et al., *DNA-Demethylating Agents Target Colorectal Cancer Cells by Inducing Viral Mimicry by Endogenous Transcripts.* Cell, 2015. **162**(5): p. 961-73.

108. Kershaw, M.H., et al., *Immunization against endogenous retroviral tumor-associated antigens.* Cancer Res, 2001. **61**(21): p. 7920-4.

109. Sacha, J.B., et al., *Vaccination with cancer- and HIV infection-associated endogenous retrotransposable elements is safe and immunogenic.* J Immunol, 2012. **189**(3): p. 1467-79.

110. Wang-Johanning, F., et al., *Human endogenous retrovirus K triggers an antigen-specific immune response in breast cancer patients.* Cancer research, 2008. **68**(14): p. 5869-5877.

111. Smith, C.C., et al., *Endogenous retroviral signatures predict immunotherapy response in clear cell renal cell carcinoma.* J Clin Invest, 2018. **128**(11): p. 4804-4820.

112. Wang-Johanning, F., et al., *Immunotherapeutic potential of anti-human endogenous retrovirus-K envelope protein antibodies in targeting breast tumors.* J Natl Cancer Inst, 2012. **104**(3): p. 189-210.

113. Wang-Johanning, F., et al., *Human endogenous retrovirus type K antibodies and mRNA as serum biomarkers of early-stage breast cancer.* Int J Cancer, 2014. **134**(3): p. 587-95.

114. Hahn, S., et al., *Serological response to human endogenous retrovirus K in melanoma patients correlates with survival probability.* AIDS Res Hum Retroviruses, 2008. **24**(5): p. 717-23.

115. Zhao, J., et al., *Expression of Human Endogenous Retrovirus Type K Envelope Protein is a Novel Candidate Prognostic Marker for Human Breast Cancer.* Genes Cancer, 2011. **2**(9): p. 914-22.

116. Gonzalez-Cao, M., et al., *Human endogenous retroviruses and cancer.* Cancer Biol Med, 2016. **13**(4): p. 483-488.

117. Bergallo, M., et al., *Transcriptional activity of human endogenous retroviruses is higher at birth in inversed correlation with gestational age.* Infect Genet Evol, 2019. **68**: p. 273-279.

118. Lee, Y.N. and P.D. Bieniasz, *Reconstitution of an infectious human endogenous retrovirus.* PLoS Pathog, 2007. **3**(1): p. e10.

119. Ono, M., et al., *Nucleotide sequence of human endogenous retrovirus genome related to the mouse mammary tumor virus genome.* J Virol, 1986. **60**(2): p. 589-98.

120. Kraus, B., et al., *Characterization of the human endogenous retrovirus K Gag protein: identification of protease cleavage sites.* Retrovirology, 2011. **8**: p. 21.

121. Chudak, C., et al., *Identification of late assembly domains of the human endogenous retrovirus-K(HML-2).* Retrovirology, 2013. **10**: p. 140.

122. Kristoffer Shalin, P.M. *Pinfish- Pipeline for de novo clustering of long transcriptomic reads*. 2020 June 10, 2021]; Available from: https://github.com/nanoporetech/pipeline-nanopore-denovo-isoforms.

123. Li, H., *Minimap2: pairwise alignment for nucleotide sequences.* Bioinformatics (Oxford, England), 2018. **34**(18): p. 3094-3100.

124. Edgar, R.C., *MUSCLE: multiple sequence alignment with high accuracy and high throughput.* Nucleic Acids Res, 2004. **32**(5): p. 1792-7.

125. Lopez-Terrada, D., et al., *Towards an international pediatric liver tumor consensus classification: proceedings of the Los Angeles COG liver tumors symposium.* Mod Pathol, 2014. **27**(3): p. 472-91.

126. Gadd, S., et al., *A Children's Oncology Group and TARGET initiative exploring the genetic landscape of Wilms tumor.* Nat Genet, 2017. **49**(10): p. 1487-1494.

127. Flockerzi, A., et al., *Expression patterns of transcribed human endogenous retrovirus HERV-K(HML-2) loci in human tissues and the need for a HERV Transcriptome Project.* BMC Genomics, 2008. **9**: p. 354.

128. Wang-Johanning, F., et al., *Expression of human endogenous retrovirus k envelope transcripts in human breast cancer.* Clinical cancer research : an official journal of the American Association for Cancer Research, 2001. **7**(6): p. 1553-1560.

129. Kleiman, A., et al., *HERV-K(HML-2) GAG/ENV antibodies as indicator for therapy effect in patients with germ cell tumors.* International journal of cancer.Journal international du cancer, 2004. **110**(3): p. 459-461.

130. Ma, W., et al., *Human Endogenous Retroviruses-K (HML-2) Expression Is Correlated with Prognosis and Progress of Hepatocellular Carcinoma.* Biomed Res Int, 2016. **2016**: p. 8201642.

131. Li, M., et al., *Downregulation of Human Endogenous Retrovirus Type K (HERV-K) Viral env RNA in Pancreatic Cancer Cells Decreases Cell Proliferation and Tumor Growth.* Clin Cancer Res, 2017. **23**(19): p. 5892-5911.

132. Rhyu, D.W., et al., *Expression of human endogenous retrovirus env genes in the blood of breast cancer patients.* Int J Mol Sci, 2014. **15**(6): p. 9173-83.

133. Grabski, D.F., et al., *Human Endogenous Retrovirus-K mRNA Expression and Genomic Alignment Data in Hepatoblastoma.* Data in Brief, 2020. **Submitted**.

134. Rooney, M.S., et al., *Molecular and genetic properties of tumors associated with local immune cytolytic activity.* Cell, 2015. **160**(1-2): p. 48-61.

135. Kassiotis, G. and J.P. Stoye, *Making a virtue of necessity: the pleiotropic role of human endogenous retroviruses in cancer.* Philos Trans R Soc Lond B Biol Sci, 2017. **372**(1732).

136. Grabski, D.F., et al., *Upregulation of human endogenous retrovirus-K (HML-2) mRNAs in hepatoblastoma: Identification of potential new immunotherapeutic targets and biomarkers.* J Pediatr Surg, 2021. **56**(2): p. 286-292.

137. Cherkasova, E., et al., *Inactivation of the von Hippel-Lindau tumor suppressor leads to selective expression of a human endogenous retrovirus in kidney cancer.* Oncogene, 2011. **30**(47): p. 4697-706.

138. Cherkasova, E., et al., *Detection of an Immunogenic HERV-E Envelope with Selective Expression in Clear Cell Kidney Cancer.* Cancer Res, 2016. **76**(8): p. 2177-85.

139. Florl, A.R., et al., *DNA methylation and expression of LINE-1 and HERV-K provirus sequences in urothelial and renal cell carcinomas.* Br J Cancer, 1999. **80**(9): p. 1312-21.

140. Singh, S., et al., *Human endogenous retrovirus K (HERV-K) rec mRNA is expressed in primary melanoma but not in benign naevi or normal skin.* Pigment Cell Melanoma Res, 2013. **26**(3): p. 426-8.

141. Löwer, R., et al., *Identification of human endogenous retroviruses with complex mRNA expression and particle formation.* Proc Natl Acad Sci U S A, 1993. **90**(10): p. 4480-4.

142. Grandi, N. and E. Tramontano, *HERV Envelope Proteins: Physiological Role and Pathogenic Potential in Cancer and Autoimmunity.* Front Microbiol, 2018. **9**: p. 462.

143. Boese, A., et al., *Human endogenous retrovirus protein cORF supports cell transformation and associates with the promyelocytic leukemia zinc finger protein.* Oncogene, 2000. **19**(38): p. 4328-36.

144. Kaufmann, S., et al., *Human endogenous retrovirus protein Rec interacts with the testicular zinc-finger protein and androgen receptor.* J Gen Virol, 2010. **91**(Pt 6): p. 1494-502.

145. Ranganathan, S., et al., *Loss of EGFR-ASAP1 signaling in metastatic and unresectable hepatoblastoma.* Sci Rep, 2016. **6**: p. 38347.

146. Bolger, A.M., M. Lohse, and B. Usadel, *Trimmomatic: a flexible trimmer for Illumina sequence data.* Bioinformatics, 2014. **30**(15): p. 2114-20.

147. Patro, R., et al., *Salmon provides fast and bias-aware quantification of transcript expression.* Nat Methods, 2017. **14**(4): p. 417-419.

148. Soneson, C., M.I. Love, and M.D. Robinson, *Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences.* F1000Res, 2015. **4**: p. 1521.

149. Love, M.I., W. Huber, and S. Anders, *Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2.* Genome Biol, 2014. **15**(12): p. 550.

150. Jornsten, R., et al., *DNA microarray data imputation and significance analysis of differential expression.* Bioinformatics, 2005. **21**(22): p. 4155-61.

151. Yu, G., et al., *clusterProfiler: an R package for comparing biological themes among gene clusters.* Omics, 2012. **16**(5): p. 284-7.

152. Wickham, H., *ggplot2: Elegant Graphics for Data Analysis*. 2016: Springer-Verlag New York.

153. Blighe K, R.S., Lewis M. *EnhancedVolcano: Publication-ready volcano plots with enhanced colouring and labeling. R package version 1.4.0*. 2019; Available from: https://github.com/kevinblighe/EnhancedVolcano.

154. Pertea, M., et al., *Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown.* Nat Protoc, 2016. **11**(9): p. 1650-67.

155. Srinivasakumar, N., et al., *The effect of viral regulatory protein expression on gene delivery by human immunodeficiency virus type 1 vectors produced in stable packaging cell lines.* J Virol, 1997. **71**(8): p. 5841-8.

156. Jacob, A.G. and C.W.J. Smith, *Intron retention as a component of regulated gene expression programs.* Human Genetics, 2017. **136**(9): p. 1043-1057.

157. Lareau, L.F., et al., *The coupling of alternative splicing and nonsense-mediated mRNA decay.* Adv Exp Med Biol, 2007. **623**: p. 190-211.

158. Black, D.L. and S.L. Zipursky, *To Cross or Not to Cross: Alternatively Spliced Forms of the Robo3 Receptor Regulate Discrete Steps in Axonal Midline Crossing.* Neuron, 2008. **58**(3): p. 297-298.

159. Doherty, J.K., et al., *The HER-2/neu receptor tyrosine kinase gene encodes a secreted autoinhibitor.* Proceedings of the National Academy of Sciences of the United States of America, 1999. **96**(19): p. 10869-10874.

160. Kaur, G., et al., *Alternative splicing of helicase-like transcription factor (Hltf): Intron retention-dependent activation of immune tolerance at the feto-maternal interface.* PLOS ONE, 2018. **13**(7): p. e0200211.

161. Li, Y., et al., *An intron with a constitutive transport element is retained in a Tap messenger RNA.* Nature, 2006. **443**(7108): p. 234-237.

162. Matsumura, M.E., et al., *Vascular injury induces posttranscriptional regulation of the Id3 gene: cloning of a novel Id3 isoform expressed during vascular lesion formation in rat and human atherosclerosis.* Arteriosclerosis, Thrombosis, and Vascular Biology, 2001. **21**(5): p. 752-758.

163. Pollard, V.W. and M.H. Malim, *The HIV-1 Rev protein.* Annu Rev Microbiol, 1998. **52**: p. 491-532.

164. Hadzopoulou-Cladaras, M., et al., *The rev (trs/art) protein of human immunodeficiency virus type 1 affects viral mRNA and protein expression via a cis-acting sequence in the env region.* J Virol, 1989. **63**(3): p. 1265-74.

165. Younis, I. and P.L. Green, *The human T-cell leukemia virus Rex protein.* Front Biosci, 2005. **10**: p. 431-45.

166. Mertz, J.A., et al., *Mouse mammary tumor virus encodes a self-regulatory RNA export protein and is a complex retrovirus.* J Virol, 2005. **79**(23): p. 14737-47.

167. Ernst, R.K., et al., *Secondary structure and mutational analysis of the Mason-Pfizer monkey virus RNA constitutive transport element.* Rna, 1997. **3**(2): p. 210-22.

168. Grüter, P., et al., *TAP, the human homolog of Mex67p, mediates CTE-dependent RNA export from the nucleus.* Mol Cell, 1998. **1**(5): p. 649-59.

169. Aibara, S., et al., *The principal mRNA nuclear export factor NXF1:NXT1 forms a symmetric binding platform that facilitates export of retroviral CTE-RNA.* Nucleic Acids Res, 2015. **43**(3): p. 1883-93.

170. Wang, B., D. Rekosh, and M.L. Hammarskjold, *Evolutionary conservation of a molecular machinery for export and expression of mRNAs with retained introns.* Rna, 2015. **21**(3): p. 426-37.

171. Kechin, A., et al., *cutPrimers: A New Tool for Accurate Cutting of Primers from Reads of Targeted Next Generation Sequencing.* J Comput Biol, 2017. **24**(11): p. 1138-1143.

172. Li, H. and R. Durbin, *Fast and accurate short read alignment with Burrows-Wheeler transform.* Bioinformatics, 2009. **25**(14): p. 1754-60.

173. Jackson, P.E.H., et al., *A novel retroviral vector system to analyze expression from mRNA with retained introns using fluorescent proteins and flow cytometry.* Sci Rep, 2019. **9**(1): p. 6467.

174. Ge, S.X., D. Jung, and R. Yao, *ShinyGO: a graphical gene-set enrichment tool for animals and plants.* Bioinformatics, 2020. **36**(8): p. 2628-2629.

175. Torrado, M., et al., *ANKRD1 specifically binds CASQ2 in heart extracts and both proteins are co-enriched in piglet cardiac Purkinje cells.* Journal of Molecular and Cellular Cardiology, 2005. **38**(2): p. 353-365.

176. Torrado, M., et al., *Intron retention generates ANKRD1 splice variants that are co-regulated with the main transcript in normal and failing myocardium.* Gene, 2009. **440**(1-2): p. 28-41.

177. Naro, C., et al., *An Orchestrated Intron Retention Program in Meiosis Controls Timely Usage of Transcripts during Germ Cell Differentiation.* Developmental Cell, 2017. **41**(1): p. 82-93.e4.

178. Hocine, S., R.H. Singer, and D. Grünwald, *RNA processing and export.* Cold Spring Harb Perspect Biol, 2010. **2**(12): p. a000752.

179. Carmody, S.R. and S.R. Wente, *mRNA nuclear export at a glance.* J Cell Sci, 2009. **122**(Pt 12): p. 1933-7.

180. Hautbergue, G.M., et al., *Mutually exclusive interactions drive handover of mRNA from export adaptors to TAP.* Proc Natl Acad Sci U S A, 2008. **105**(13): p. 5154-9.

181. Herold, A., L. Teixeira, and E. Izaurralde, *Genome-wide analysis of nuclear mRNA export pathways in Drosophila.* Embo j, 2003. **22**(10): p. 2472-83.

182. Hieronymus, H. and P.A. Silver, *Genome-wide analysis of RNA-protein interactions illustrates specificity of the mRNA export machinery.* Nat Genet, 2003. **33**(2): p. 155-61.

183. Katahira, J., et al., *The Mex67p-mediated nuclear mRNA export pathway is conserved from yeast to human.* Embo j, 1999. **18**(9): p. 2593-609.

184. Viphakone, N., et al., *TREX exposes the RNA-binding domain of Nxf1 to enable mRNA export.* Nat Commun, 2012. **3**: p. 1006.

185. Rodrigues, J.P., et al., *REF proteins mediate the export of spliced and unspliced mRNAs from the nucleus.* Proc Natl Acad Sci U S A, 2001. **98**(3): p. 1030-5.

186. Huang, Y., T.A. Yario, and J.A. Steitz, *A molecular link between SR protein dephosphorylation and mRNA export.* Proc Natl Acad Sci U S A, 2004. **101**(26): p. 9666-70.

187. Müller-McNicoll, M., et al., *SR proteins are NXF1 adaptors that link alternative RNA processing to mRNA export.* Genes Dev, 2016. **30**(5): p. 553-66.

188. Wang, K., et al., *Intronless mRNAs transit through nuclear speckles to gain export competence.* J Cell Biol, 2018. **217**(11): p. 3912-3929.

189. Zuckerman, B., et al., *Gene Architecture and Sequence Composition Underpin Selective Dependency of Nuclear Export of Long RNAs on NXF1 and the TREX Complex.* Mol Cell, 2020.

190. Ben-Yishay, R., et al., *Imaging within single NPCs reveals NXF1's role in mRNA export on the cytoplasmic side of the pore.* J Cell Biol, 2019. **218**(9): p. 2962-2981.

191. Katz, Y., et al., *Analysis and design of RNA sequencing experiments for identifying isoform regulation.* Nature Methods, 2010. **7**(12): p. 1009-1015.

192. Shen, S., et al., *rMATS: Robust and flexible detection of differential alternative splicing from replicate RNA-Seq data.* Proceedings of the National Academy of Sciences of the United States of America, 2014. **111**(51): p. E5593-E5601.

193. Forrest, S.T., et al., *Intron Retention Generates a Novel Id3 Isoform That Inhibits Vascular Lesion Formation.* Journal of Biological Chemistry, 2004. **279**(31): p. 32897-32903.

194. Bray, M., et al., *A small element from the Mason-Pfizer monkey virus genome makes human immunodeficiency virus type 1 expression and replication Rev-independent.* Proceedings of the National Academy of Sciences of the United States of America, 1994. **91**(4): p. 1256-1260.

195. Borden, K.L.B., *The Nuclear Pore Complex and mRNA Export in Cancer.* Cancers (Basel), 2020. **13**(1).

196. Bharathavikru, R., et al., *Transcription factor Wilms' tumor 1 regulates developmental RNAs through 3' UTR interaction.* Genes Dev, 2017. **31**(4): p. 347-352.

197. Pardo, B., et al., *Homologous recombination and Mus81 promote replication completion in response to replication fork blockage.* EMBO Rep, 2020. **21**(7): p. e49367.

198. Gatfield, D. and E. Izaurralde, *REF1/Aly and the additional exon junction complex proteins are dispensable for nuclear mRNA export.* J Cell Biol, 2002. **159**(4): p. 579-88.

199. Björk, P., J.O. Persson, and L. Wieslander, *Intranuclear binding in space and time of exon junction complex and NXF1 to premRNPs/mRNPs in vivo.* J Cell Biol, 2015. **211**(1): p. 63-75.

200. Le Hir, H., et al., *The spliceosome deposits multiple proteins 20-24 nucleotides upstream of mRNA exon-exon junctions.* Embo j, 2000. **19**(24): p. 6860-9.

201. Oesterreich, F.C., et al., *Splicing of Nascent RNA Coincides with Intron Exit from RNA Polymerase II.* Cell, 2016. **165**(2): p. 372-381.

202. Chen, L.L. and G.G. Carmichael, *Altered nuclear retention of mRNAs containing inverted repeats in human embryonic stem cells: functional role of a nuclear noncoding RNA.* Mol Cell, 2009. **35**(4): p. 467-78.

203. Coyle, J.H., et al., *The Tpr protein regulates export of mRNAs with retained introns that traffic through the Nxf1 pathway.* Rna, 2011. **17**(7): p. 1344-56.

204. Galy, V., et al., *Nuclear retention of unspliced mRNAs in yeast is mediated by perinuclear Mlp1.* Cell, 2004. **116**(1): p. 63-73.

205. Bonnet, A., H. Bretes, and B. Palancade, *Nuclear pore components affect distinct stages of intron-containing gene expression.* Nucleic Acids Res, 2015. **43**(8): p. 4249-61.

206. Saroufim, M.A., et al., *The nuclear basket mediates perinuclear mRNA scanning in budding yeast.* J Cell Biol, 2015. **211**(6): p. 1131-40.

207. Sloan, E.A., et al., *Limited nucleotide changes in the Rev response element (RRE) during HIV-1 infection alter overall Rev-RRE activity and Rev multimerization.* J Virol, 2013. **87**(20): p. 11173-86.

208. Graham, F.L. and A.J. van der Eb, *A new technique for the assay of infectivity of human adenovirus 5 DNA.* Virology, 1973. **52**(2): p. 456-67.

209. Jin, L., et al., *Tap and NXT promote translation of unspliced mRNA.* Genes Dev, 2003. **17**(24): p. 3075-86.

210. Love, M.I., C. Soneson, and R. Patro, *Swimming downstream: statistical analysis of differential transcript usage following Salmon quantification.* F1000Res, 2018. **7**: p. 952.

211. Rivera, M.N. and D.A. Haber, *Wilms' tumour: connecting tumorigenesis and organ development in the kidney.* Nat Rev Cancer, 2005. **5**(9): p. 699-712.

212. Call, K.M., et al., *Isolation and characterization of a zinc finger polypeptide gene at the human chromosome 11 Wilms' tumor locus.* Cell, 1990. **60**(3): p. 509-20.

213. Gessler, M., et al., *Homozygous deletion in Wilms tumours of a zinc-finger gene identified by chromosome jumping.* Nature, 1990. **343**(6260): p. 774-8.

214. Huff, V., *Wilms' tumours: about tumour suppressor genes, an oncogene and a chameleon gene.* Nat Rev Cancer, 2011. **11**(2): p. 111-21.

215. Davies, R.C., E. Bratt, and N.D. Hastie, *Did nucleotides or amino acids drive evolutionary conservation of the WT1 +/-KTS alternative splice?* Hum Mol Genet, 2000. **9**(8): p. 1177-83.

216. Barbaux, S., et al., *Donor splice-site mutations in WT1 are responsible for Frasier syndrome.* Nat Genet, 1997. **17**(4): p. 467-70.

217. Hohenstein, P. and N.D. Hastie, *The many facets of the Wilms' tumour gene, WT1.* Hum Mol Genet, 2006. **15 Spec No 2**: p. R196-201.

218. Motamedi, F.J., et al., *WT1 controls antagonistic FGF and BMP-pSMAD pathways in early renal progenitors.* Nat Commun, 2014. **5**: p. 4444.

219. Bor, Y.C., et al., *The Wilms' tumor 1 (WT1) gene (+KTS isoform) functions with a CTE to enhance translation from an unspliced RNA with a retained intron.* Genes Dev, 2006. **20**(12): p. 1597-608.

220. Thäte, C., C. Englert, and M. Gessler, *Analysis of WT1 target gene expression in stably transfected cell lines.* Oncogene, 1998. **17**(10): p. 1287-94.

221. Lee, T.H. and J. Pelletier, *Functional characterization of WT1 binding sites within the human vitamin D receptor gene promoter.* Physiol Genomics, 2001. **7**(2): p. 187-200.

222. Vajjhala, P.R., et al., *The Wilms' tumour suppressor protein, WT1, undergoes CRM1-independent nucleocytoplasmic shuttling.* FEBS Lett, 2003. **554**(1-2): p. 143-8.

223. Laity, J.H., H.J. Dyson, and P.E. Wright, *Molecular basis for modulation of biological function by alternate splicing of the Wilms' tumor suppressor protein.* Proc Natl Acad Sci U S A, 2000. **97**(22): p. 11932-5.

224. Larsson, S.H., et al., *Subnuclear localization of WT1 in splicing or transcription factor domains is regulated by alternative splicing.* Cell, 1995. **81**(3): p. 391-401.

225. Rauscher, F.J., 3rd, et al., *Binding of the Wilms' tumor locus zinc finger protein to the EGR-1 consensus sequence.* Science, 1990. **250**(4985): p. 1259-62.

226. Stoll, R., et al., *Structure of the Wilms tumor suppressor protein zinc finger domain bound to DNA.* J Mol Biol, 2007. **372**(5): p. 1227-45.

227. Nishikawa, T., et al., *RNA Binding by the KTS Splice Variants of Wilms' Tumor Suppressor Protein WT1.* Biochemistry, 2020. **59**(40): p. 3889-3901.

228. Nowicka, M. and M.D. Robinson, *DRIMSeq: a Dirichlet-multinomial framework for multivariate count outcomes in genomics.* F1000Res, 2016. **5**: p. 1356.

229. Benassayag, C., et al., *Human c-Myc isoforms differentially regulate cell growth and apoptosis in Drosophila melanogaster.* Mol Cell Biol, 2005. **25**(22): p. 9897-909.

230. Kretzner, L., E.M. Blackwood, and R.N. Eisenman, *Myc and Max proteins possess distinct transcriptional activities.* Nature, 1992. **359**(6394): p. 426-9.

231. Niksic, M., et al., *The Wilms' tumour protein (WT1) shuttles between nucleus and cytoplasm and is present in functional polysomes.* Hum Mol Genet, 2004. **13**(4): p. 463-71.

232. Hammes, A., et al., *Two splice variants of the Wilms' tumor 1 gene have distinct functions during sex determination and nephron formation.* Cell, 2001. **106**(3): p. 319-29.

233. Kann, M., et al., *Genome-Wide Analysis of Wilms' Tumor 1-Controlled Gene Expression in Podocytes Reveals Key Regulatory Mechanisms.* J Am Soc Nephrol, 2015. **26**(9): p. 2097-104.

234. Lefebvre, J., et al., *Alternatively spliced isoforms of WT1 control podocyte-specific gene expression.* Kidney Int, 2015. **88**(2): p. 321-31.

235. Dong, L., et al., *Integration of Cistromic and Transcriptomic Analyses Identifies Nphs2, Mafb, and Magi2 as Wilms' Tumor 1 Target Genes in Podocyte Differentiation and Maintenance.* J Am Soc Nephrol, 2015. **26**(9): p. 2118-28.

236. Toska, E. and S.G. Roberts, *Mechanisms of transcriptional regulation by WT1 (Wilms' tumour 1).* Biochem J, 2014. **461**(1): p. 15-32.

237. Potluri, S., et al., *Isoform-specific and signaling-dependent propagation of acute myeloid leukemia by Wilms tumor 1.* Cell Rep, 2021. **35**(3): p. 109010.

238. Zhang, Y., et al., *The role of WT1 in breast cancer: clinical implications, biological effects and molecular mechanism.* Int J Biol Sci, 2020. **16**(8): p. 1474-1480.

239. Nishida, S. and H. Sugiyama, *Immunotherapy Targeting WT1: Designing a Protocol for WT1 Peptide-Based Cancer Vaccine.* Methods Mol Biol, 2016. **1467**: p. 221-32.

240. Davies, R.C., et al., *WT1 interacts with the splicing factor U2AF65 in an isoform-dependent manner and can be incorporated into spliceosomes.* Genes Dev, 1998. **12**(20): p. 3217-25.

241. Caricasole, A., et al., *RNA binding by the Wilms tumor suppressor zinc finger proteins.* Proc Natl Acad Sci U S A, 1996. **93**(15): p. 7562-6.

242. Kalkat, M., et al., *MYC Deregulation in Primary Human Cancers.* Genes (Basel), 2017. **8**(6).

243. Amati, B., et al., *Oncogenic activity of the c-Myc protein requires dimerization with Max.* Cell, 1993. **72**(2): p. 233-45.

244. Blackwood, E.M. and R.N. Eisenman, *Max: a helix-loop-helix zipper protein that forms a sequence-specific DNA-binding complex with Myc.* Science, 1991. **251**(4998): p. 1211-7.

245. Sabò, A., M. Doni, and B. Amati, *SUMOylation of Myc-family proteins.* PLoS One, 2014. **9**(3): p. e91072.

246. Kalkat, M., et al., *MYC Protein Interactome Profiling Reveals Functionally Distinct Regions that Cooperate to Drive Tumorigenesis.* Mol Cell, 2018. **72**(5): p. 836-848.e7.

247. Blackwood, E.M., et al., *Functional analysis of the AUG- and CUG-initiated forms of the c-Myc protein.* Mol Biol Cell, 1994. **5**(5): p. 597-609.

248. Udtha, M., et al., *Upregulation of c-MYC in WT1-mutant tumors: assessment of WT1 putative transcriptional targets using cDNA microarray expression profiling of genetically defined Wilms' tumors.* Oncogene, 2003. **22**(24): p. 3821-6.

249. Han, Y., et al., *Transcriptional activation of c-myc proto-oncogene by WT1 protein.* Oncogene, 2004. **23**(41): p. 6933-41.

250. Wu, C., et al., *WT1 enhances proliferation and impedes apoptosis in KRAS mutant NSCLC via targeting cMyc.* Cell Physiol Biochem, 2015. **35**(2): p. 647-62.

251. Ladomery, M.R., et al., *Presence of WT1, the Wilm's tumor suppressor gene product, in nuclear poly(A)(+) ribonucleoprotein.* J Biol Chem, 1999. **274**(51): p. 36520-6.