Mechanisms and impact of Post-Transcriptional Exon Shuffling (PTES)

Ginikachukwu Osagie Izuogu

Doctor of Philosophy

Institute of Genetic Medicine

Newcastle University

September 2016

ABSTRACT

Most eukaryotic genes undergo splicing to remove introns and join exons sequentially to produce protein-coding or non-coding transcripts. Post-transcriptional Exon Shuffling (PTES) describes a new class of RNA molecules, characterized by exon order different from the underlying genomic context. PTES can result in linear and circular RNA (circRNA) molecules and enhance the complexity of transcriptomes.

Prior to my studies, I developed PTESFinder, a computational tool for PTES identification from high-throughput RNAseq data. As various sources of artefacts (including pseudogenes, template-switching and others) can confound PTES identification, I first assessed the effectiveness of filters within PTESFinder devised to systematically exclude artefacts. When compared to 4 published methods, PTESFinder achieves the highest specificity (~0.99) and comparable sensitivity (~0.85). To define sub-cellular distribution of PTES, I performed *in silico* analyses of data from various cellular compartments and revealed diverse populations of PTES in nuclei and enrichment in cytosol of various cell lines. Identification of PTES from chromatin-associated RNAseq data and an assessment of co-transcriptional splicing, established that PTES may occur during transcription. To assess if PTES contribute to the proteome, I analyzed sucrose-gradient fractionated data from HEK293, treated with arsenite to induce translational arrest and dislodge ribosomes. My results showed no effect of arsenite treatment on ribosome occupancy within PTES transcripts, indicating that these transcripts are not generally bound by polysomes and do not contribute to the proteome.

To investigate the impact of differential degradation on expression levels of linear and circRNAs, I analyzed the PTES population within RNAseq data of anucleate cells and established that most PTES transcripts are circular and are enriched in platelets 17-to-188-fold relative to nucleated tissues. For some genes, only reads from circRNA exons were detectable, suggesting that platelets have lost >90% of their progenitor mRNAs, consistent with time-dependent degradation of platelets transcriptomes. However, some circRNAs exhibit read density patterns suggestive of miRNA induced degradation.

Finally, a linear PTES from RMST locus has been implicated in pluripotency maintenance using limited RNAseq data from human embryonic stem cells (hESC). To identify other PTES transcripts with similar expression patterns, I analyzed RNAseq data from H9 ESC differentiation series. Statistical analyses of PTES transcripts identified during cellular differentiation established that PTES expression changes track with that of cognate linear transcripts and accumulate upon differentiation. Contrary to previous reports, the dominant transcript from RMST is circular and increases in abundance during differentiation. Functional

analyses demonstrating the role of RMST in pluripotency maintenance had targeted exons within the predicted circRNA, suggesting previously unreported functional relevance for circRNAs.

Declaration

I, Osagie Izuogu, declare that no material documented in this thesis has been submitted in support of a degree or other qualification in this or any other university. This thesis represents my own work and any collaborative work is acknowledged where necessary.

Osagie Izuogu

Dedication

This thesis is dedicated to Denise, my wife; Tiana, Dante and Luca, my children. Thank you all for your unwavering support and understanding throughout my research. I love you all.

Acknowledgements

Enormous gratitude to Dr. Mike Jackson and Dr. Mauro Santibanez-Koref for the opportunity to undertake my research under their supervision. I am particularly grateful for your guidance, support and patience, and for being available for all formal and informal discussions. I sincerely look forward to future opportunities to work with you, thank you. To Dr. Alhassan, thank you for your guidance and assistance with *in vitro* analyses, I appreciate your help.

To Prof. David Elliott and members of Elliott's lab, thank you for your support and for providing answers to my numerous questions. I would like to thank current and former members of the statistical genetics group (IGM), especially, Dr. Miossecc, Dr. Griffin, Dr. Ayers, Dr. Xu, Dr. Ainsworth and Dr. Howey. I thank you all for your words of advice, encouragement and impromptu tutorials. Our informal meetings helped shape my thoughts.

Special thanks to my collaborators - Prof Lako (Newcastle University, UK), Dr. Ghevaert & colleagues (Cambridge University, UK) - and my panel members - Prof. Heather Cordell, Dr. Andreas Werner and Dr. Venables; I thoroughly enjoyed working on this project and my research benefitted from your expertise, words of encouragement and guidance, thank you. To Dr. Harry Mountain (Staffordshire University, UK) and Ada Izuogu (University of Toledo, Ohio), thank you both for proofreading my thesis and your valuable comments.

I am grateful to my family and friends, for their unconditional love and support throughout my study, thank you all. I acknowledge the Biotechnology and Biological Sciences Research Council for funding my research. Ultimately, I thank God for his mercies.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

ABBREVIATIONS	FULL MEANING
AR	Abundance Ratio
BED	Browser Extensible Data
BLAST	Basic Local Alignment Search Tool
BLAT	BLAST Like Alignment Tool
bp	base pairs
CAGE	Cap Analysis of Gene Expression
cDNA	Complementary DNA
circRNA	Circular Ribonucleic Acid
CSR	Co-transcriptional Splicing Rate
DE	Differential Expression
DNA	Deoxyribonucleic Acid
dsRNA	double stranded Ribonucleic Acid
ENCODE	Encyclopedia of DNA Elements
ERCC	External RNA Controls Consortium
ESC	Embryonic Stem Cells
EST	Expressed Sequence Tags
FPKM	Fragments Per Kilobase per Million
IGM	Institute of Genetic Medicine
IGV	Integrated Genomics Viewer
Indel	Insertion and deletions
J2SE	Java 2 Standard Edition
JPM	Junctions Per Million
JSpan	Junction Span
LincRNA	Long intergenic non-coding Ribonucleic Acid
LINES	Long Interspersed Nuclear Elements
miRNA	Micro Ribonucleic Acid
mRNA	Messenger RNA
NCBI	National Centre for Biotechnology Information
ncRNA	Non-coding Ribonucleic Acid
NGS	Next Generation Sequencing
NMD	Nonsense Mediated Decay
NUMTS	Nuclear Mitochondria DNA Sequences
ORF	Open Reading Frame
PE	Paired-end
PCR	Polymerase Chain Reaction
PID	Percent Identity

PTES	Post-Transcriptional Exon Shuffling
qPCR	Quantitative Polymerase Chain Reaction
RACE	Rapid Amplification of cDNA Ends
RDBMS	Relational Database Management System
REFSEQ	Reference Sequence
RNA	Ribonucleic Acid
RNA-SEQ	RNA Sequencing
RPKM	Reads Per Kilobase per Million
rRNA	Ribosomal RNA
RT-PCR	Reverse Transcriptase Polymerase Chain Reaction
SAGE	Serial Analysis of Gene Expression
SAM	Sequence Alignment Map
SINES	Short Interspersed Nuclear Elements
SL	Splice Leader
snoRNA	Small Nucleolar Ribonucleic Acid
snRNP	Small Nuclear Ribonucleic Particles
SRA	Sequence Read Archive
TAE	tris-acetate-EDTA
UCSC	University of California Santa Cruz

CHAPTER 1. Introduction

1.1 Transcriptome Diversity in Humans

Transcriptome profiling by deep sequencing of Ribonucleic acid (RNA) has aided the discovery of various RNA species and improved our understanding of their diverse functions. Before the availability of high-throughput sequencing technologies, RNAs were mainly considered as intermediates in the information flow between DNA and proteins. In the last decade, it has become clear that eukaryotic transcriptomes are more complex and diverse than previously thought. We now understand that, although only about 2% of the human genome contributes to the proteome, >75% of the genome is transcribed (Birney et al. 2007; Djebali et al. 2012). GENCODE, a catalog of mouse and human transcripts, currently (version 19) has a total of 198,619 human transcripts, and less than half of these are annotated protein-coding transcripts (Table 1, (Harrow et al. 2012)). Since its first release in 2009, new transcripts are being discovered and annotated at the rate of >10,000 transcripts annually, the majority of which are non-protein coding and have no known functional relevance.

RNA Class	Genes	Transcripts
Protein Coding	19797	79795
Non-Protein Coding		
Ribosomal RNA (rRNA)	544	544
Micro RNA (miRNA)	4093	4093
Small nuclear RNA (snRNA)	1896	1896
Small nucleolar RNA (snoRNA)	949	961
Small cajal RNA (scaRNA)	49	49
Long intergenic non-coding RNA (lincRNA)	7678	13301
others	25492	97980
Total:	**60498**	**198619**

Table 1.1. **Summary of GENCODE annotated transcripts**. Various RNA species within GENCODE v. 19 annotated transcripts (Harrow et al., 2012)

Transcription of both strands of the genome is also more extensive than previously thought (Katayama et al. 2005; Sanna et al. 2008). According to the FANTOM consortium, >20% of human transcripts have antisense pairs (transcripts from opposite strand) (Katayama et al. 2005; Sanna et al. 2008; Chen et al. 2004). From a locus, transcription of both strands can produce overlapping transcripts, sometimes resulting in multiple isoforms of both transcripts. Some small non-coding RNAs (ncRNAs) are embedded within introns of genes and are cleaved out post-transcription (Wilusz et al. 2009; Tran & Hutvagner 2013). For instance, *BAALC*, a gene highly expressed in acute myeloid leukaemia patients (Tanner et al. 2001), has two RefSeq

annotated isoforms, is flanked by two antisense transcripts (*BAALC-AS1* and *BAALC-AS2*) and has *MIR3151*, a micro RNA (miRNA) embedded within its first intron (Fig 1.1).

One direct result of increased transcriptional output of any locus is competition for both RNA-binding proteins and RNA-RNA interactions via Watson-Crick base pairing (Sen et al. 2014). MiRNAs are short (~22 bp) ncRNAs that can regulate protein-coding transcripts by inducing degradation and reducing translation of transcripts harbouring their binding sites (Ha & Kim 2014; Tran & Hutvagner 2013). Because of sequence identity, ncRNAs transcribed from the same genomic space may compete with mRNAs for miRNA binding resulting in increased expression of the mRNAs, as in the case of *PTENP1* and *PTEN* (Poliseno et al. 2010). Similarly, ncRNAs antisense to mRNAs, can mask binding sites for miRNAs, resulting in increased stability, as in the case of *BACE1-AS* and *BACE1* (Faghihi et al. 2008; Faghihi et al. 2010). Because each miRNA can target over 200 mRNAs (Kapranov et al. 2007), these ncRNAs can also act as competing endogenous RNAs (ceRNAs) *in trans*, impacting the expression of mRNAs not transcribed from the same locus (Kapranov et al., 2007). *LincRNA-RoR* (Long intergenic non-coding RNA, regulator of reprogramming) has been shown to sequester miRNAs that act on *OCT4, SOX2* and *NANOG*, known transcription factors that promote pluripotency in human embryonic stem cells (HESC)(Wang et al. 2013).

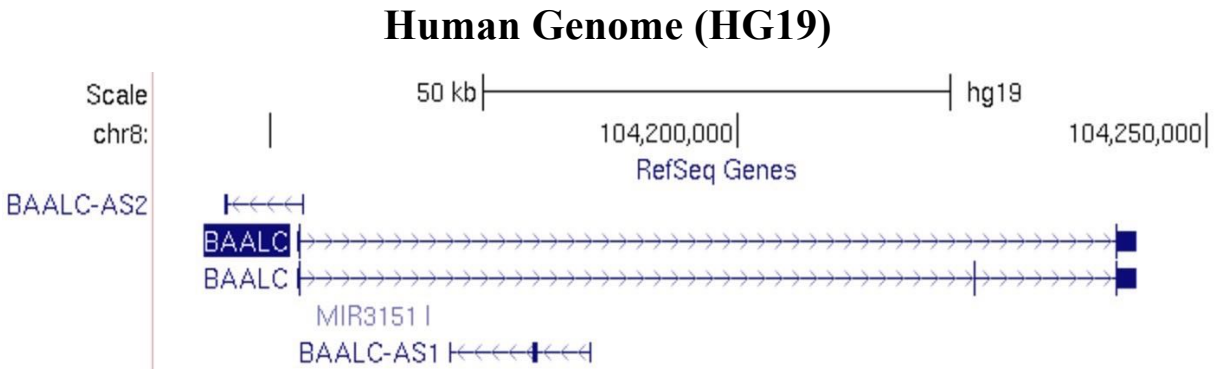


Figure 1.1: **Bi-directional Transcription**. RefSeq annotations of two *BAALC* isoforms, one consisting of 2 exons, the other 3 exons. Two antisense transcripts are also produced from the same genomic space. MIR3151 is embedded within intron 1 of *BAALC* (Pruitt et al. 2007). Coordinates are based on the human genome (HG19).

Post-Transcriptional Exon Shuffling (PTES) describes a new class of RNA molecules characterized by exon order different from the underlying genomic context (Al-Balool et al. 2011) and is the subject of this thesis. Transcripts arising from PTES can be linear or circular and originate from the same genomic positions as both protein-coding and non-coding transcripts, thus, enhancing the complexity of eukaryotic transcriptomes. In subsequent sections of this chapter, I first briefly introduce mechanisms driving transcriptome diversity in

eukaryotes, including PTES. Existing knowledge about mechanisms such as alternative splicing, trans-splicing and other mechanisms resulting in chimeric transcripts are relevant to characterizing PTES. After this synopsis, I then introduce current methods for PTES identification from RNA extracts and factors affecting *in silico* PTES identification.

1.2 Major Sources of Transcriptome Diversity

1.2.1 Alternative Splicing

Most eukaryotic genes consist of exonic and intronic sequence regions. Introns are removed by the spliceosome, a complex of proteins and small nuclear RNAs (U1, U2, U4, U5 and U6), before translation (Elliott & Ladomery, 2011). Splicing was discovered in the late 1970s by two research groups - Berget and Sharp (1977) and Chow et al. (1977). They observed size variations between pre-mRNA transcripts in the nucleus and their mature counterparts in the cytoplasm (Berget et al. 1977; Chow et al. 1977). Using a method called R-looping, which allows for the hybridization of RNA and DNA in high concentrations of formamide, mature transcripts of an adenovirus gene were observed to hybridize to short segments of its DNA, leaving gaps corresponding to intronic sequences.

For efficient splicing reactions in multi-cellular organisms with large introns, exons are first recognized by binding of serine-arginine (SR) proteins to short sequence segments within exons and introns: Exonic and Intronic Splicing Enhancers (ESE and ISE respectively) (Faustino & Cooper 2003; Wang et al. 2004). Equivalently, splicing repressor proteins can bind silencer sequences within exons and introns (ESS and ISS), to inhibit exon recognition and splicing (Faustino and Cooper 2003; Wang et al. 2004). Exon-Intron boundaries are also recognized by conserved sequence motifs that allow for spliceosome assembly. Together with enhancers and repressors, these motifs constitute the splice code necessary for efficient removal of introns (Fig. 1.2). Typically, splicing involves two trans-esterification reactions. Firstly, U1 snRNP binds to the donor splice site (characterized by CAG|GU consensus sequence) and U2 binds to a conserved branch point sequence within the intron to be excised. The association of U2 with the branch point causes a bulge of the conserved adenosine residue, facilitating a chemical attack on the donor splice site, resulting in a loop structure called lariat intermediate and a 2'-5' phosphodiester bond. Synchronously, U5 binds to the acceptor splice site, U4 and U6 assemble and associate with U5, stabilizing the spliceosome on the transcript. In the second stage, the first nucleotide of the intron at the acceptor splice site is chemically attacked by the 3'-OH of the donor exon resulting in the release of the lariat intermediate and joining of the

two exons. Some eukaryotic genes contain introns with different splice site consensus sequences (characterized by AT-AC at intron termini), and are processed by a minor spliceosome (Elliott & Landomery, 2011). Splicing of these introns involve U11, U12, U5, U4atac and U6atac) and are very rare.

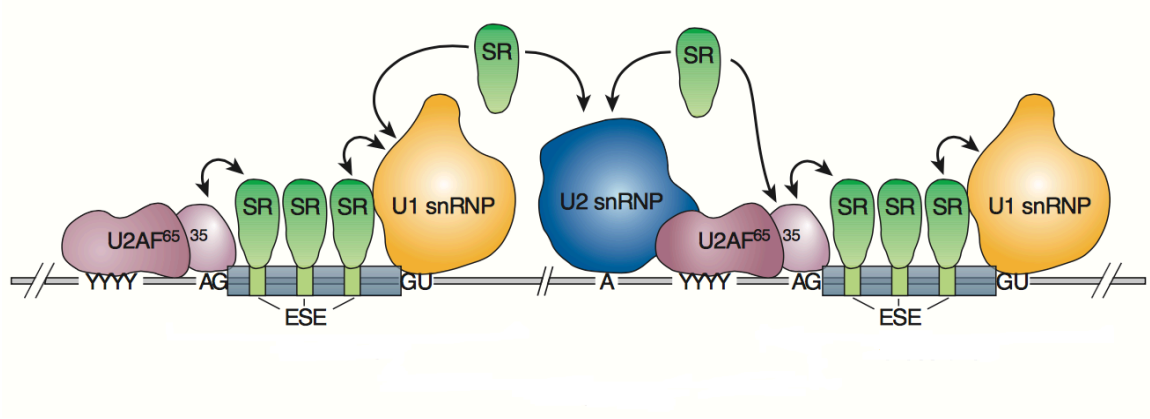


Figure 1.2: **Splice site recognition**. Consensus sequence around exon-intron junctions (GU—AG) are recognized by spliceosomal proteins for splicing to occur. RNA recognition motifs within exons called exonic sequence enhancers (ESE) are bound by serine-argenine (SR) proteins to aid exon definition. Other sequence motifs within introns (including: Conserved branch point adenosine and pyrimidine tract [YYYY]) are also bound by proteins to promote splicing. Adapted from (Maniatis & Tasic 2002).

During splicing, alternative events can occur. Exons can be skipped, alternative 5' or 3' splice sites can be used, introns can be retained and alternative promoters or termination sites can be utilized (for reviews, see (Keren et al. 2010; Xing & Lee 2006)). All of these events can result in diverse isoforms of transcripts from the same locus. Importantly, various factors can influence the occurrence of these alternative events. One such factor is the rate of transcription elongation at the locus. Recent reports suggest that most modifications to pre-mRNAs occur co-transcriptionally (Ameur et al. 2011; Bentley 2014; Girard et al. 2012; Merkhofer et al. 2014). It is estimated that the elongation rate of RNA polymerase II is between 1-4 kb/min and splicing can occur over several minutes (Bentley, 2014). As splice sites become available to be processed, spliceosomal proteins are assembled on a first come first served basis (de la Mata et al. 2010). Splice sites can compete for these proteins. In some cases, exons with weak splice signals can be skipped as a result of an increased elongation rate making available other splice sites that provide competition for spliceosomal proteins (Bentley, 2014). Conversely, transcription of large introns can cause delays in elongation, resulting in inclusion of neighboring exons.

1.2.2 Transcription of 'Junk DNA'

Transcriptome diversity is also enhanced by excised introns, transcripts from intergenic and non-coding gene regions - genomic regions once thought to be 'junk DNA' or 'genomic dark matter' (Wilusz et al. 2009; Clark et al. 2013; Kapranov et al. 2010) or transcriptionally silent (Birney et al., 2007).

Intronic transcripts: Following splicing, lariat intermediates (containing introns) are linearised by debranching enzymes and rapidly degraded (Ruskin & Green 1985). Typically, introns have low half-lives of between 5 and 7 minutes (Bentley 2014). In some cases, however, discarded introns are protected from degradation or are only partially degraded (Qian et al. 1992; Yin et al. 2012; Zhang et al. 2014). Introns containing templates for miRNAs and snoRNAs are protected from exonuclease activity by the proteins involved in their respective biogenesis pathways. For miRNAs, two main pathways for biogenesis exist: one involving direct transcription of primary miRNA (pri-miRNA), which is capped and polyadenylated; a second pathway for miRNAs embedded within introns (mirtrons) of mRNAs involves splicing and debranching (Westholm & Lai 2011; Ha & Kim 2014; Tran & Hutvagner 2013). In both cases, the pri-miRNA is further processed in the nucleus by a micro-processor complex consisting of *DGCR8* and *DROSHA*. Cleaving of the pri-miRNA produces pre-miRNAs of 70 - 100bp in length, prior to export to the cytoplasm and further processing by another endonuclease, *DICER*, into mature miRNAs (Fig 1.3). Biogenesis of snoRNAs embedded within introns is also thought to involve nucleases (Filipowicz & Pogacic 2002; Rearick et al. 2011). A recent study reported the identification of 19 transcripts consisting of partially degraded introns with snoRNAs at both termini (Yin et al., 2012; Zhang et al., 2014). These transcripts, termed sno-lncRNAs, were also found in various cell lines and other primates, suggesting conservation across species. One mouse-specific sno-lncRNA consists of *Snord33* and *Snord34* (Zhang et al., 2014). The biogenesis of this sno-lncRNA is thought to involve alternative splicing, as both snoRNAs are embedded within adjacent introns of *RPL13A*. Notably, the host intron of *Snord34* is sometimes retained (Hubbard et al. 2002), further enhancing transcript diversity at this locus.

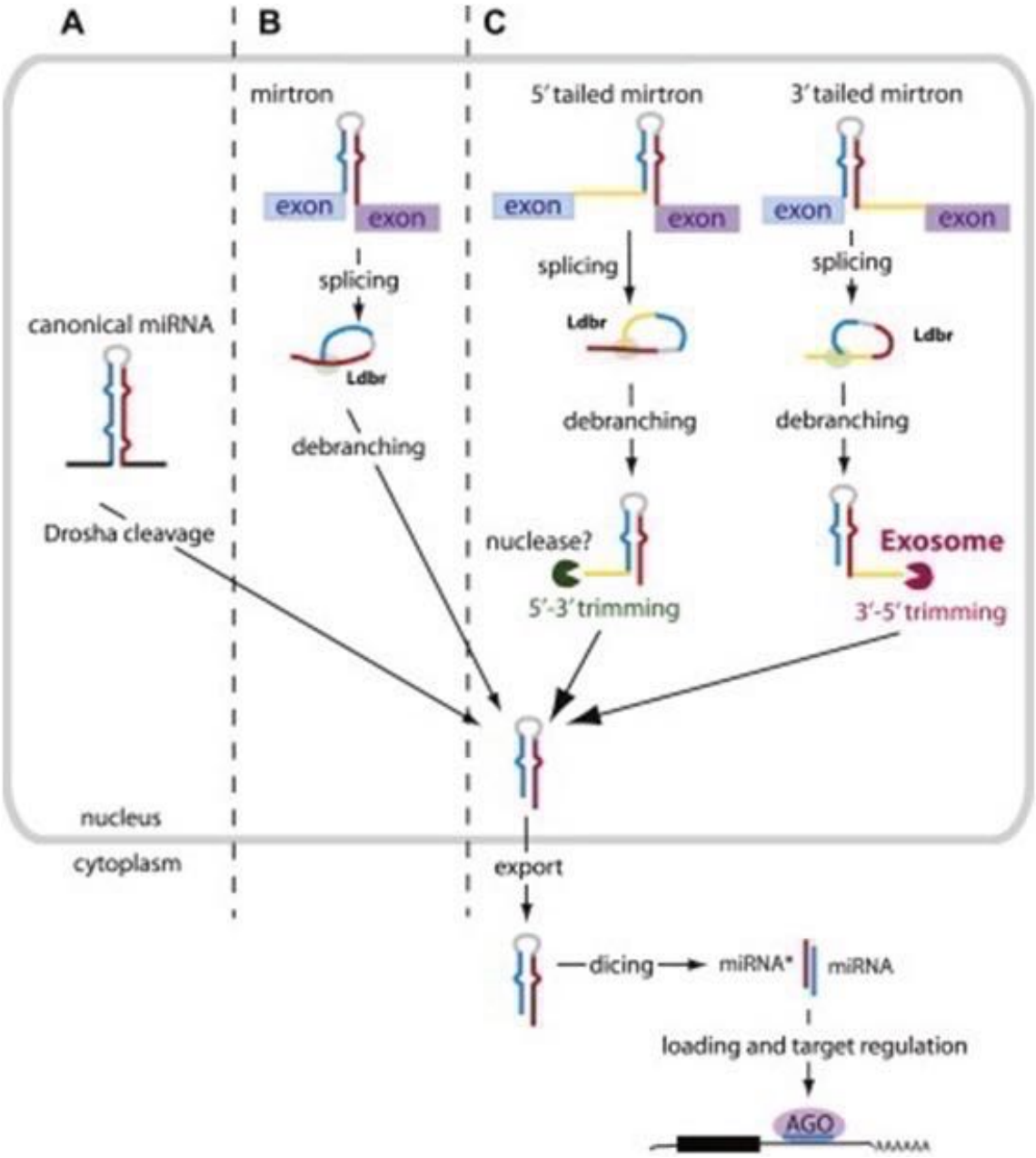


Figure 1.3. **Overview of MiRNA biogenesis mechanisms.** A) MiRNA host genes are transcribed into primary miRNA (pri-miRNA), processed by *DROSHA* prior to translocation to the cytoplasm B) Some miRNAs are embedded in introns of other genes (mirtron) and are cleaved out into pri-miRNA, following splicing and debranching C) and exonuclease trimming of flanking sequence. Figure from Westholm & Lai, (2011).

Pseudogenes: New genes and gene families can arise through duplications of DNA segments. Duplications can occur when chromosomes are mis-aligned during meiosis or as a result of dissociation and re-attachment of DNA polymerase to the template strand during replication (Zhang 2003). Genes resulting from duplications diverge, acquiring new functions and structure through mutations and re-arrangements. A class of pseudogenes can arise from mutations to duplicated genes which affects their transcription, splicing or translation into functional proteins (Mighell et al. 2000; Pink & Carter 2013). These pseudogenes typically retain the

structure of their parental genes, have high sequence identity with their parental genes and can contain introns (Mighell et al., 2000). If transcribed, splicing and translation may occur. Within this class of pseudogenes are unitary pseudogenes (Zhang et al. 2010), which do not have any functional parental genes and do not originate from recent duplication events. Zhang et al., (2010) reported the identification of 11 unitary pseudogenes, 2 of which appear to be pseudogenes in non-human primates but have functional and non-functional alleles within the human population. Another class of pseudogenes arise from retrotransposition. Mature transcripts are sometimes reverse transcribed and integrated into the genome (Pei et al. 2012; Kalyana-Sundaram et al. 2012; Kaessmann et al. 2009). As a result, these pseudogenes are referred to as processed pseudogenes because they lack introns and contain stretches of adenosine residues. In rare cases, these processed pseudogenes are integrated near a functional promoter that drives their transcription and can result in proteins (Kim et al. 2014; Ji et al. 2015; Xu & Zhang 2016).

Natural antisense transcripts: Natural antisense transcripts (NATs) were first discovered over 35 years ago (Lacatena & Cesareni 1981), in the bacterial plasmid, ColE1 (Hershfield et al. 1974; Naito & Uchida 1980). The copy number and replication of ColE1 was found to be controlled by a sense-antisense pair of transcripts (Lacatena & Cesareni 1981; Wagner & Simons 1994). The sense transcript initiates replication by forming an RNA-DNA hybrid with the plasmid DNA. This hybrid is cleaved by RNase H, resulting in fragments that prime DNA synthesis (Wagner & Simons, 1994). However, the transcription of the antisense transcript antagonises replication by forming a duplex with the sense transcript, subsequently impacting plasmid copy number (Wagner & Simons, 1994). Since then, thousands of NATs have been identified and are typically characterized by low abundance, in some cases, >10 fold lower than that of their sense counterparts (Pelechano & Steinmetz 2013). NATs can broadly be classified into three groups - Convergent, Divergent and Internal, depending on their orientation relative to their sense transcripts (Pelechano & Steinmetz 2013; Wight & Werner 2013). Convergent NATs have tail-to-tail orientations to their sense counterpart, with their 3' termini overlapping that of sense counterparts. Conversely, divergent NATs overlap with sense transcripts at 5' termini, whilst some NATs are completely overlapped (internal) by their sense counterparts. Antisense transcripts can originate from cryptic promoters, bidirectional promoters or promoters independent from those used by their sense counterparts (Pelecheno & Steinmetz, 2013). Because they can form duplexes with other transcripts as a result of sequence complementarity at overlapped regions, NATs can mask binding sites for RNA binding proteins and miRNAs (Pelecheno & Steinmetz, 2013, Wight & Werner, 2013). This is exemplified by

the sense-antisense transcripts from the *BACE1* locus (Faghihi et al., 2008 & 2010), which form a duplex that masks miRNA binding sites, resulting in increase in stability of the sense mRNA. Additionally, NATs have been shown to induce epigenetic changes, affecting the transcription of genes *in cis* and *in trans* (Pelecheno & Steinmetz, 2013). A classic example is *Tsix*, antisense to X inactivation specific transcript (*Xist*), which is expressed in inactive X chromosomes. *Xist* has been shown to recruit the polycomb repressive complex (*PRC2*) to promote the formation of heterochromatin and inactivation of the chromosome (Pontier & Gribnau 2011; Pelechano & Steinmetz 2013). However, when expressed, *Tsix* antagonises the expression of *Xist*, by facilitating epigenetic changes at its promoter region (Pontier & Gribnau, 2011).

1.2.3 Chimeric Transcripts

Chimeric transcripts refer to transcripts consisting of segments from more than one RNA molecule. Various mechanisms can result in chimeric RNAs and are briefly introduced below:
Fused genes: Fused genes can arise from chromosome translocations, where fragments from two chromosomes emerge from breaks and are reattached incorrectly, with each chromosome receiving the others fragment. Although some translocations can result in stable chromosomes, such phenomena, in some cases, denote cancers and can result in chimeric transcripts with deleterious effects. Genes can be disrupted as a result of chromosome beaks, reattachment to another gene on a different chromosome results in fused genes. Gene fusion in some cases disrupts the open reading frame resulting in rapid degradation. However, in rare cases, hybrid proteins are produced. A classic example is the reciprocal translocation between chromosomes 9 and 22 that characterizes chronic myelogenous leukaemia. This translocation results in a transcribed fused gene, *BCR-ABL*, subsequently producing at least two variant hybrid proteins (Lichty et al. 1998; Ren 2005).

Another phenomenon is the ongoing translocation and integration of chloroplasts and mitochondrial DNA fragments into nuclear genomes. In humans, over 700 regions of the genome consist of nuclear-mitochondria DNA sequences (NUMTs). These are regions where fragments of mitochondria DNA have been inserted into the genome, presumably during chromosomal breaks (Ricchetti et al. 2004; Lenglez et al. 2010). In an investigation of human NUMTs, 23 out of 27 human-specific NUMTs were found within genes, mostly in introns, but can result in chimeric RNA transcripts. Some NUMTs are transcribed (D. Wang et al. 2014) and can result in protein. The *humanin* protein for instance, is encoded in the mitochondria, however 13 humanin-like loci have been identified on various chromosomes in the nuclear genome (Bodzioch et al. 2009), and there is evidence of a protein product 99% identical to *humanin* (Tajima et al. 2002).

Trans-Splicing: Trans-splicing - unlike cis-splicing which occurs within single RNA molecules - occurs between 2 RNA molecules. In trans-splicing, one RNA molecule provides the donor splice site (along with preceding RNA sequence in the template), which is then joined to the acceptor splice site of a different RNA molecule and a Y-shaped intermediate is discarded (Murphy et al. 1986). This phenomenon is common in unicellular organisms and is well characterized in nematodes (Davis et al. 1995; Lei et al. 2016). There are two types of trans-splicing, homotypic and heterotypic trans-splicing (Takahara et al. 2000; Takahara et al. 2002).
Homotypic trans-splicing: This is the splicing together of two RNA molecules from the same locus to produce a chimeric transcript. A typical example is the *lola* gene in Drosophila, which has been shown to extensively undergo trans-splicing to produce transcripts with complex exon arrangements not explainable by alternative splicing or genomic re-arrangements (Horiuchi & Aigaki 2006).
Heterotypic trans-splicing: This is the splicing together of two RNA molecules from different loci. Many unicellular organisms undergo polycistronic transcription, where multiple overlapping genes on the same strand are produced in one transcriptional unit (Salgado et al. 2000). Transcripts within these transcriptional units are further processed by the addition of a common untranslated exon, known as a splice leader. The splice leader sequence is typically transcribed from a highly expressed gene, provides a cap structure to aid translation but does not contribute to proteome diversity.

Read-Through Transcripts: Transcription is initiated at the promoter region, upstream from the gene to be transcribed and is signalled to terminate downstream from the gene. Two non-overlapping genes on the same strand are separated by an intergenic non-coding region. However, due to weak transcription signals, transcription can proceed beyond a gene into another gene before termination, producing a chimeric transcript (Kapranov et al. 2007; Kannan et al. 2011). Splicing of this multi-locus transcript can result in exons of both genes within the same transcript (Akiva et al. 2006). In fact, reports show that the vast majority of these transcripts join the penultimate exon of the first gene to the second exon of the second gene, removing the last exon of the first gene and the first exon of the second gene as part of the discarded lariat intermediate (Nacu et al. 2011). As reported in Akiva et al., (2006), novel exons within the intergenic region are sometimes included in the processed chimeric transcript. As both heterotypic trans-splicing events and read-through transcription involve splicing between two transcripts, the inclusion of novel exons is a distinguishing feature of read-through transcripts. Proximity between parental genes can also be used to distinguish between both

mechanisms. Notably, a common feature of 2369 chimeric transcripts identified by Kannan et al., (2011), is the small genomic distance (median: ~2KB) between parental genes, suggesting read-through transcription.

Typically, read-through transcripts and heterotypic trans-spliced transcripts in mammals are expressed at low levels and appear to be tissue-specific (Frenkel-Morgenstern et al. 2012). Parental genes of most of these transcripts have higher than average expression patterns suggesting that the production of these chimeric transcripts is potentially unregulated. In some cases, however, chimeric transcripts can be highly expressed and characterize disease stage. For instance, a chimeric transcript (from *SLC45A3* and *ELK4* loci) is highly expressed in prostate cancer patients (Kannan et al., 2011). Fused genes originating from chromosomal rearrangements also denote various cancers. One fate of chimeric transcripts is that they are rapidly degraded due to the alteration of an open reading frame (ORF) (Akiva et al., 2006; Nacu et al., 2011). Transcripts not subjected to nonsense mediated decay (NMD) can be translated into hybrid proteins with more than one protein domain. Notably, a heterotypic trans-spliced transcript consisting of 5' exons of *JAZF1* gene and 3' exons of *JJAZI* gene has been shown to be translated in endometrial stroma cells (Li et al. 2008).

1.3 Post-transcriptional Exon Shuffling (PTES)

Many annotated transcripts do not have known functional relevance and many transcripts remain unannotated, even undiscovered. Recently, a novel class of transcripts has been described; and can be produced from the same genomic location as both protein coding and non-coding transcripts, further enhancing the complexity of eukaryotic transcriptomes. Post-transcriptional Exon Shuffling (PTES) describes the existence of transcripts with re-arranged exon order different from the order in the genome (Al-Balool et al. 2011). In the absence of exon duplication or genomic re-arrangement, a 5-exon gene, for instance, can be spliced to produce an un-rearranged (canonical) transcript: 1-2-3-4-5 (Fig 1.4). In PTES, splicing can occur between a downstream donor splice site and an upstream acceptor site, resulting in a re-arranged linear transcript with repeated exons or a backsplice joining a subset of exons in a circular RNA (circRNA) (Fig 1.4). Importantly, PTES events can result in either linear or circular transcripts and are characterized at the sequence level by a non-canonical exon-exon junction. For simplicity, throughout this thesis, the rearranged exon-exon junction characterising PTES is termed a PTES junction; PTES transcripts are inferred from the junction with predicted internal exons for circRNA and repeated exons for linear PTES. For instance, the inferred transcript for a back-splice between exons 5 and 2 is assumed to include exons 3

and 4 as internal exons. Where a linear PTES is inferred, all exons (including repeated exons) are predicted to constitute the transcript, as depicted in Fig 1.4.

In PTES, rearrangements are typically intragenic, unlike in heterotypic trans-splicing and read-through transcription. Similarly, circRNAs resulting from PTES differ from previously observed viroid RNAs and undebranched lariat intermediates covalently joined at 2' - 5' (Zhang et al. 2013; Jeck & Sharpless 2014). Viroids are plant pathogens, which self-replicate within hosts to produce rolling circle RNAs that are cleaved at sites flanked by ~ 9bp repeat sequences (Sanger et al. 1976). Cleavage and ligation reactions in viroids are not mediated by the spliceosome.

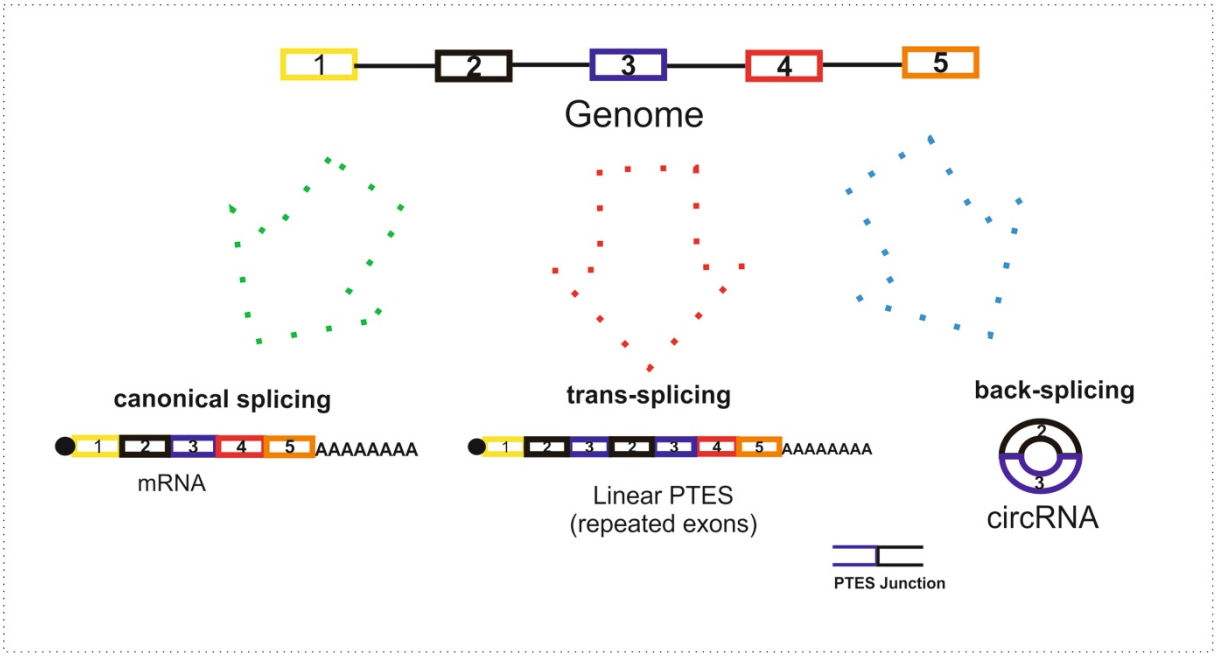


Figure 1.4. **Post-transcriptional Exon Shuffling (PTES).** Schematic diagram of various transcripts likely from a hypothetical 5-exon gene. From a gene, canonical splicing occurs to produce mRNA transcripts; trans-splicing of two linear molecules from the gene can result in linear PTES with repeated exons. Back-splicing of a downstream splice site to an upstream splice site can result in circular RNA. In both cases, the distinguishing feature of PTES is the rearranged exon-exon junction.

The earliest observation of PTES was reported in Nigro et al., (1991). The authors serendipitously observed 'scrambled exons' within a transcript while trying to characterize the exon arrangement of the *DCC* (deleted in colorectal cancer) gene. Without prior knowledge of exon arrangement, primer pairs spanning each exon pairing in both directions were used in amplification reactions and only the amplicon in the correct orientation expected. To their surprise, they observed 2 amplicons for a single exon-exon pairing, indicating the presence of both the un-rearranged (canonical) transcript and a rearranged transcript. Comparing expression estimates of both transcripts, they estimated that the rearranged transcript was expressed at 0.1% of that of the canonical transcript (Nigro et al. 1991). Three additional PTES transcripts

were identified in both human cells and rat tissues, with the highest expression observed in rat brain. Similar observations were made in *ETS-1*, an oncogene (Cocquerelle et al. 1992; Cocquerelle et al. 1993; Bailleul 1996). PTES transcripts identified in these studies lacked 5' cap structures and polyA tails, suggesting circularity. Capel et al., (1993) showed that PTES transcripts produced from the *Sry* locus of adult rats were circular by digesting RNA extracts with RNase H (an enzyme that preferentially degrades RNA-DNA hybrids), such that when probed by sequence fragment complementary to circularized exon, the number of fragments observed was used to deduce circularity (see 1.3.3 for details).

Although earlier PTES observations were of circRNAs, later observations of linear PTES transcripts produced by homotypic trans-splicing were also reported. Takahara et al. (2000 & 2002) reported the identification of linear PTES transcripts emanating from *Sp1*, a transcription factor. Upon amplification of the expected canonical transcript, the authors observed other amplicons of larger sizes. Follow up RNase H digestion and sequencing confirmed these amplicons to be both linear and polyadenylated. Similarly, Al-Balool et al. (2011) reported that some PTES transcripts were enriched in polyA+ RNA fractions relative to total RNA fractions and could be amplified from PTES defining junction into untranslated regions, confirming the presence of exons not expected in analogous circular transcripts. Many other studies have made similar observations and concluded that linear PTES transcripts arise from splicing between 2 RNA molecules of the same transcript (Caudevilla et al. 1998; Caldas et al. 1998; Frantz et al. 1999; Rigatti et al. 2004; Dixon et al. 2007).

The overriding notion after these discoveries was that PTES transcripts were potentially products of aberrant splicing (Cocquerelle et al. 1993; Salzman et al. 2012). This notion was supported by their very low expression relative to canonical transcripts. However, thousands of these transcripts have now been identified from various mammalian cell lines and tissues. PTES events have been observed in plants (Lu et al. 2015; Ye et al. 2015; P. L. Wang et al. 2014), archaea (Danan et al. 2012), fungi (Wang et al., 2014), flies (Dixon et al. 2005; Ashwal-Fluss et al. 2014; Westholm et al. 2014), worms (Memczak et al. 2013; Ivanov et al. 2014) and rodents (Rigatti et al. 2004; Rybak-Wolf et al. 2015; Zaphiropoulos 1997), suggesting that their existence may not solely be due to aberrant splicing. These transcripts also differ from other chimeric transcripts that are lowly expressed but originate from highly expressed parental genes. PTES events can result in highly expressed transcripts, with expression patterns comparable to canonical linear transcripts from the same loci (Salzman et al. 2012; Starke et al. 2014; Al-Balool et al. 2011; Capel et al. 1993). Capel et al., (1993) reported that the circRNA

from the *Sry* locus exists in adult rat in the absence of the canonical transcript, indicating that this circRNA is the dominant transcript from that locus.

1.3.1 Mechanisms of PTES Formation

Progress has been made in elucidating the mechanisms of PTES formation but our understanding of how they are regulated remains poor. Two mechanisms of PTES biogenesis have been proposed (Fig 1.5), one involving pairing of introns flanking PTES exons and the other involving re-splicing of skipped exons within discarded lariat intermediates.

Flanking intron sequence pairing: Mechanistically, production of linear PTES transcripts is thought to include intron pairing (Dixon et al., 2007). Inverted complementary sequence repeats within flanking introns are thought to pair during transcription, bringing splice sites of two RNAs together to be spliced. Trans-splicing mediated by intron pairing has been demonstrated *in vivo* (Takahara et al. 2005) and *in vitro* (Solnick 1985). In Takahara et al., (2005), long introns were shown to promote trans-splicing of *Sp1* minigene constructs in the HepG2 cell line. In the same study, the authors found that RNA polymerase II pause sites within flanking introns of PTES exons promote intron pairing and subsequent trans-splicing. In Solnick (1985), two minigene constructs - each consisting of one exon from adenovirus and one exon from human beta-globin gene, flanked by introns with inverted repeats - produced trans-spliced transcripts containing either exons from adenovirus or beta-globin exons.

Intron pairing can also result in circRNAs. Following the identification of circRNA from the *Sry* locus (circSry) by Capel et al., (1993), Dubin et al., (1995) demonstrated that circSry could be produced *in vitro*. The rat *Sry* locus consists of a single exon flanked by ~15 kb of inverted repeat sequence that facilitates pairing of sequence around the splice sites, resulting in non-canonical splicing and circularization (Dubin et al. 1995). Inverted repeats within introns can include transposable elements such as *Alu* repeats (Jeck et al. 2013). A recent report by Liang & Wilusz (2014) showed that as little as 30 - 40 bp of flanking sequence complementarity is required for circularisation. Restricting their experiments to circRNAs produced from *ZKSCAN1*, *HIPK3* and *EPHB4*, the authors progressively reduced the size of flanking intron *Alu* repeats. They also determined that, in some cases, stronger intron pairing may inhibit circularisation (Liang & Wilusz 2014). Similar results were obtained when complementary inverted *DNAREP1_DM* repeats flanking circRNA transcripts from *lacasse2* gene in drosophila were progressively shortened (Kramer et al. 2015).

Re-splicing within lariat intermediates: Re-splicing of processed transcripts is a rare event but has been observed in transcripts of two genes within cancer cells (Chen et al. 2015; Kameyama et al. 2012). Aberrant spliced transcripts were identified from the tumor susceptibility gene 101 (*TSG101*) where cryptic splice sites within exons 2 and 9 were utilised in splicing of the mature mRNA (Kameyama et al., 2012). It has been theorised that further splicing may occur within discarded lariat intermediates prior to debranching (Salzman et al., 2012; Jeck et al., 2013). Splice sites of exons skipped following alternative splicing can be brought together in close proximity for backsplicing to occur. This premise is based on the observation that some circRNAs constitute skipped exons and that their expression correlates with the expression of linear isoforms lacking those exons (Surono et al. 1999). Various studies have reported a link between PTES formation and alternative splicing (Wilusz 2015; Kelly et al. 2015), but the extent of this link is still being explored. Zaphiropoulos (1997) identified circRNAs from the cytochrome P450 locus, consisting of exons absent in various canonically spliced isoforms. Similarly, whilst investigating circRNAs predicted from observed canonical isoforms of the dystrophin gene, Surono et al. (1999) identified 12 circRNAs consisting of skipped exons but failed to identify 3, 2 of which were single exons known to be alternatively spliced. Despite these findings, there are reports of circRNAs involving exons not alternatively spliced (Surono et al. 1999), suggesting that this mechanism for PTES formation is not universal. A classic example is the circRNA from the single exon gene, *Sry*. Consistent with this premise, Jeck et al., 2013 observed that only 45% of circRNAs identified from human fibroblasts had a detectable alternatively spliced linear transcript lacking backspliced exons.

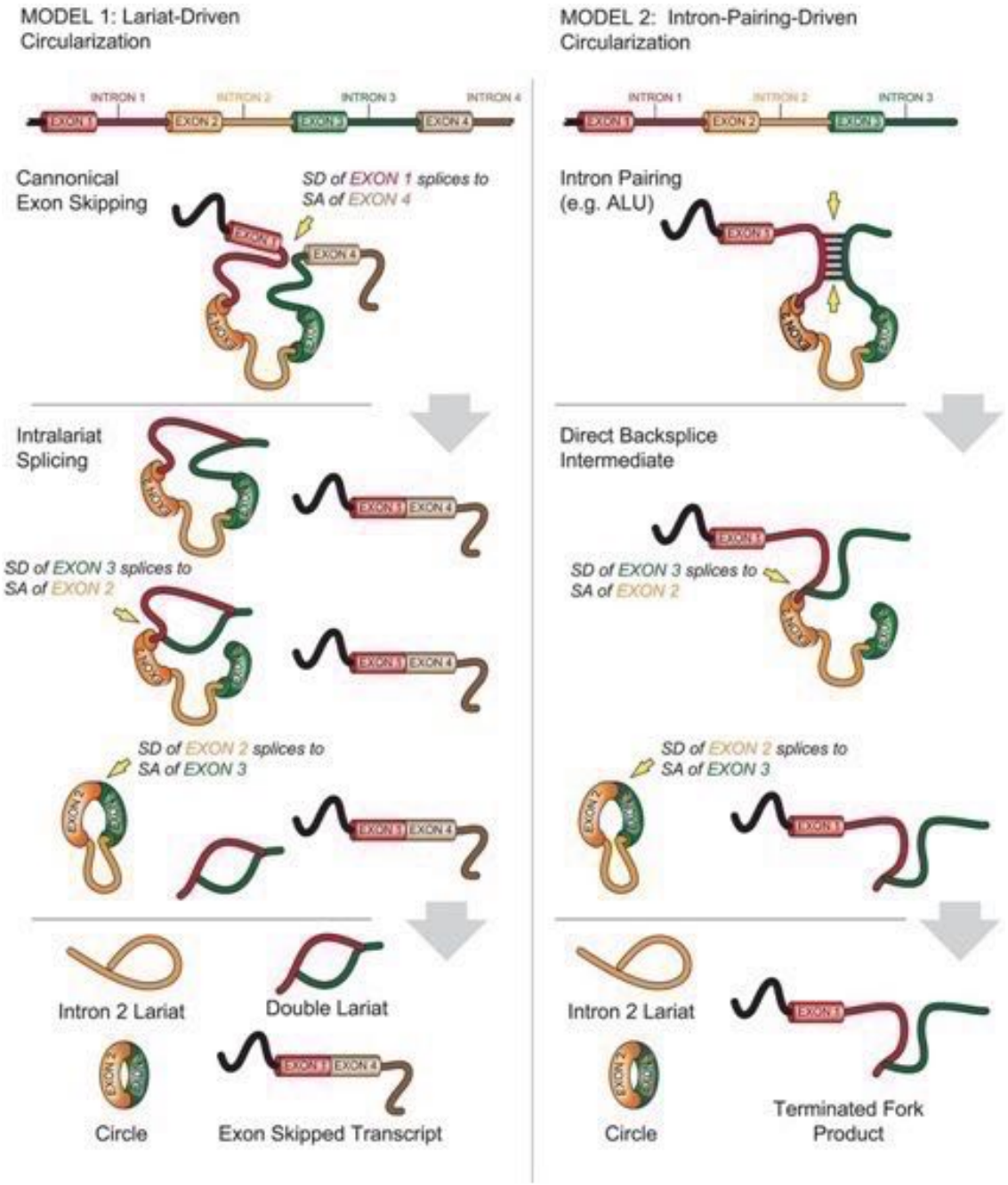


Figure 1.5. Two models of PTES biogenesis. **Left)** Re-splicing of skipped exons within lariat intermediates can result in backsplice. **Right)** Intron-pairing mediated by complementary repeat sequences can bring distal splice sites in close proximity to be spliced in tandem resulting in PTES. Figure taken from Jeck et al., 2013.

1.3.2 PTES Formation is regulated by RNA Binding Proteins (RBP)

Several lines of evidence have established that the vast majority of PTES events occur at precise exon-intron junctions, indicating the role of the spliceosome in PTES biogenesis (Ashwal-Fluss et al., 2014; Jeck & Sharpless, 2014; Starke et al., 2015). For instance, inhibiting the assembly of U4, U5 and U6 snRNPs by treating cells with isoginkgetin, abolished PTES (Starke et al., 2014). In another study, mutating known splice sites from GU to CA, diminished circRNA biogenesis in *PVT1* and *CRKL* loci (Ashwal-Fluss et al., 2014). Various RNA binding

proteins (RBPs) enhance or repress splicing by 1) aiding recognition and use of splice sites; 2) by recruiting other spliceosomal proteins and 3) by promoting RNA secondary structures that can affect the choice of splice sites. Recently, two splice factors were shown to facilitate PTES biogenesis by promoting secondary structures favorable for circularization. First, Muscleblind (*MBNL*) was shown to aid the production of circRNA from exon 2 of its locus (Ashwal-Fluss et al., 2014). The observation that circRNA from *MBNL is* abundant in Drosophila fly heads but low in Drosophila S2 expressing a different isoform of the gene, led Ashwal-Fluss and colleagues to theorize the role of *MBNL* in PTES. They transfected Drosophila S2 cells with 3 variants of *MBNL* and found one variant to increase circularization by 13 fold. This was also accompanied by a 2-fold reduction in the levels of canonically spliced transcripts from *MBNL*, suggesting a relationship between PTES and mRNA expression in this locus, where the expression level of *MBNL* directly regulates PTES production from its locus (Ashwal-Fluss et al., 2014). This observation was followed by the identification of *MBNL* binding sites within introns flanking the circularized exon. It is thought that, even for exons flanked by short introns, thus lacking flanking inverted repeats, *MBNL* binding may facilitate intron pairing resulting in PTES. In another study, Conn et al., (2015) demonstrated that Quaking (*QKI* - a splice factor), acts to regulate biogenesis of up to 33% of PTES transcripts. *QKI* binding sites within introns flanking exons involved in PTES were mutated in that study, resulting in reduced PTES abundance. Furthermore, modulating expression levels of *QKI* induced changes in PTES abundance, further demonstrating the role of *QKI* in PTES biogenesis (Conn et al. 2015).

Splice factors can induce or repress PTES biogenesis by affecting the choice of splice sites. To identify trans-acting factors that may regulate circularization, Kramer et al., (2015) knocked down various RBPs involved in splicing, including 3 serine-arginine (SR) proteins: *SF2*, *SRp54* and *SRSF6*. All 3 SR proteins independently induced >2X increase in abundance of circRNA for *laccasse2* gene in drosophila, suggesting repressive effects of these splice factors. In contrast, knock down of *Hrb867F*, a heterogenous ribonucleoprotein particle, resulted in decrease of circRNA abundance (Kramer et al., 2015). Furthermore, a group of enzymes, called adenosine deaminases acting on RNA (*ADAR*) was recently shown to negatively regulate PTES (Ivanov et al., 2014). *ADAR* proteins have various functions within cells, one of which involves preventing the formation of double stranded RNA (dsRNA) (Ota et al. 2013). In the nucleus, pre-mRNA transcripts are edited by *ADAR1,* by replacing the amino group on adenosine with an oxygen atom, resulting in inosine. Editing of adenosine residues to inosine weakens dsRNA, as inosine weakly bonds with thymine, and occurs at higher frequency in *Alu* repeats within introns (Elliott & Landomery, 2011). Inverted complementary *Alu* repeats have been shown to promote intron pairing and PTES (Jeck et al., 2013; Zhang et al., 2014), and *ADAR1* activity

inhibits PTES events. This was demonstrated in Ivanov et al., (2014), where the authors reported an increase in circRNA abundance upon *ADAR1* knockdown.

1.3.3 In vitro Methods for PTES identification

Early observations of viroid circular RNAs were made using electron microscopes (Sanger et al., 1976). Since then, various molecular biology protocols have been adapted to identify PTES events from RNA extracts or enrich for circRNA transcripts. The distinguishing feature of PTES at the sequence level is the rearranged exon-exon junction. This singular feature is crucial to all approaches for *in vitro* PTES identification. These approaches can broadly be grouped into 3:

1. Amplification approach: RNA extracts are typically reverse transcribed into complementary DNA (cDNA) to be used as templates for amplification using polymerase chain reaction (PCR). To identify PTES, primers are designed to specifically amplify region spanning the PTES junction. For instance, to detect a hypothetical PTES involving exons 3 and 2, the forward primer is designed to be homologous to position within exon 3 and the reverse primer within exon 2, ensuring that an amplicon of the expected size spans the PTES junction (depicted in Figure 1.6). As conventional PCR is not quantitative, probes can also be designed to span across the PTES junction in quantitative PCR (qPCR) assays.

This approach however, can be prone to technical variability in a reverse transcriptase (RT) dependent manner (Yu et al. 2014); and can be a major source of artefacts in PTES detection (more on this in section 1.3.5). The use of 2 different types of RT (specifically MMLV-derived and AMV-derived) in cDNA synthesis has been proposed as a way to confirm the validity of PTES and reduce the likelihood of PCR artefacts (Yu et al., 2014).

Studies have inferred linearity and circularity of specific PTES transcripts based on the RNA fraction screened (Al-Balool et al. 2011; Jeck et al. 2013; Wu et al. 2013). For instance, as linear PTES transcripts are presumed to be polyadenylated, identifying PTES enriched in polyA+ or cap-selected RNA fractions may suggest linearity. Similarly, PTES from non-polyadenylated RNA may provide initial evidence of circularity (Salzman et al., 2012).

2. Hybridization approach: To avoid RT-PCR based assays and associated artefacts, hybridization methods (such as northern blotting) are commonly used to detect PTES from RNA extracts. In northern blotting, labeled oligonucleotide probes homologous to the PTES junction sequence are first synthesized and then mixed with RNA immobilized on a filter membrane. Hybridization occurs by base pairing between probe and target sequence, if the

sample contains the PTES of interest. For detection, probes not hybridized are washed off and the filter typically visualized on X-ray film.

This approach can be used to define the full transcript size and exon-intron structure of PTES of interest, presenting an advantage over RT-PCR. However, unsatisfactory results and potential false predictions may result from low specificity of probes used. Additionally, although probes can be generated synthetically, PCR amplicons can also be used as probes and can result in PCR artefacts. This approach has been used to experimentally validate various PTES transcripts (Hansen et al. 2011; Memczak et al. 2013; Salzman et al. 2012; Salzman et al. 2013).

Combining hybridization of DNA probe with RNase H treatment can be used to distinguish between circRNA and linear molecules (Capel et al., 1993). RNase H is an endonuclease that preferentially cleaves RNA-DNA hybrids. For instance, to establish the structure of a hypothetical PTES between exons 3 and 2, DNA probe homologous to either exon can be used in hybridization. Following cleavage by RNase H to deplete the exon targeted, 1, 2 or 3 fragments will be observed upon visualization, to indicate the detection of a circRNA, linear canonical transcript or linear PTES respectively. As a result of repeated exons in the hypothetical linear PTES, three fragments are expected upon cleavage. Similarly, a probe designed to hybridize re-arranged exons, spanning PTES junction, will form a substrate for RNase H, producing a large fragment for undigested linear transcripts, a single shorter fragment for circRNA or 2 short fragments for linear PTES (depicted in Figure 1.6).

3. Enrichment approach: As some PTES events are rare, it may be desirable to enrich for specific RNA molecules. Ribosomal RNAs (rRNAs) are known to be the most abundant RNA species in total RNA extracts (Choy et al. 2015). To enrich for other RNA species (including PTES transcripts), methods have been devised for depleting levels of rRNAs (Adiconis et al. 2013). These methods start by binding rRNAs using RNA probes and then removing bound rRNAs, enriching other RNA species. To specifically enrich for linear polyadenylated transcripts, oligo-dT selection is typically performed. For circular RNAs, RNase R, 3’ -> 5’ exonuclease is routinely used to degrade linear molecules in RNA extracts, enriching for circRNAs. This approach can be combined with RT-PCR, northern blotting or *in silico* identification methods to specifically identify circRNAs (Jeck et al., 2013).

In Hansen et al. (2011), another circRNA enrichment method was utilised. Total RNA extracts were treated with tobacco acid phosphatase to remove cap-structures of linear molecules, before degrading with 5’ -> 3’ exonucleases and enriching circRNAs (Hansen et al., 2011; Jeck & Sharpless, 2014). Enrichment methods can however, result in variation in

expression estimates between replicates, depending on sensitivity and concentration of enzymes used in enrichment.

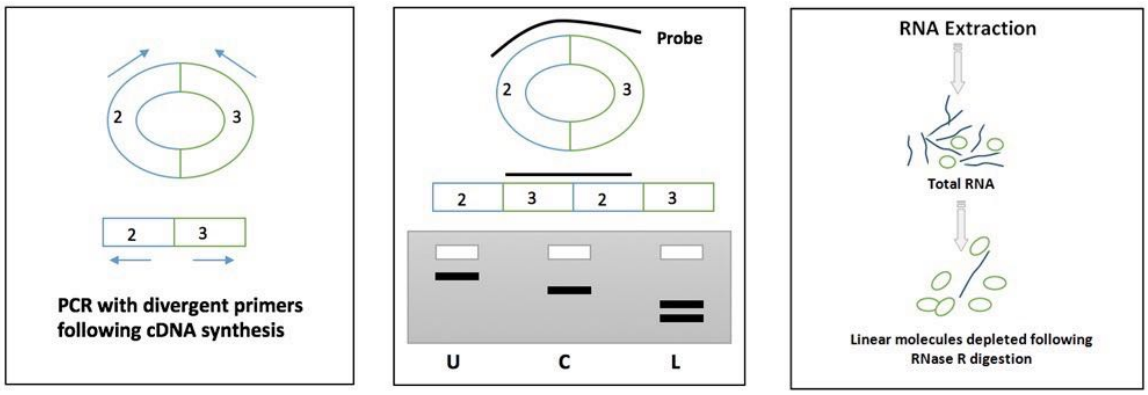


Figure 1.6. **Examples of *in vitro* PTES identification methods. Left)** PTES junction can be amplified using divergent PCR primers, spanning junction. **Middle)** Circularity or linearity of PTES junction can be inferred by probing with oligonucleotides homologous with PTES junction and digesting with RNase H, an endonuclease that preferentially degrades RNA-DNA hybrids. Linear canonical transcripts will be undigested (U), and appear larger when visualized. CircRNAs (C) with the probed PTES junction will show a single smaller fragment, due to the removal of region cleaved and the absence of some exons. Two bands will be observed for linear PTES (L) with the probed junction. **Right)** Enrichment of circular RNA molecules is typically achieved by treating total RNA with RNase R, an exonuclease that degrades linear molecules.

1.3.4 Approaches to in silico PTES identification

Prior to high throughput RNA sequencing, methods employed for transcript identification and quantification were either: 1) laborious and low-throughput *in vitro* protocols, such as qPCR; 2) hybridization-based Microarrays or 3) low-throughput sequencing of short tags from expressed transcripts. Microarrays utilize probes attached to glass slides or chips. Fluorescently labeled RNA extracts are hybridized to these probes to detect the presence of expressed transcripts and quantified by estimating fluorescence intensity. Microarray chips specific to exons and exon-junctions can be generated to profile the expression of these features (Sanchez-Pla et al. 2012). Apart from the obvious limitation that experiments utilizing microarrays were hypothesis driven and knowledge of transcript sequence was required to generate probes, this methodology also suffered from cross-hybridization of probes and limited dynamic range for estimating transcript abundance (Zhao et al. 2014). Methods relying on sequencing of short tags - such as Cap analysis gene expression (CAGE) and Serial analysis of gene expression (SAGE) (Tuteja & Tuteja 2004) - do not require prior knowledge of transcript sequence but only generated data from short (9 -13 bp) regions of transcripts (Tuteja and Tuteja 2004). In the case of CAGE, short sequence tags around the 5' cap structures are generated and used to estimate gene expression (Kodzius et al. 2006). Discriminating between isoforms of the same gene is a

challenge with these methods, as isoforms can share the same sequence from which a tag was derived.

RNA sequencing (RNAseq) addresses many of the limitations of earlier methods. RNAseq experiments start by isolating or enriching RNA species of interest, mainly by size separation, polyA+ capture, ribosomal RNA depletion or digestion by exonuclease - to name a few. cDNA fragmentation follows synthesis to reduce templates to sizes required by respective sequence platforms. As RNAseq requires low amounts of starting template, templates may be amplified before sequencing, in a platform-dependent manner (Mardis 2008; Margulies et al. 2005; Sanchez-Pla et al. 2012). Post-sequencing, sequence reads can be assembled to re-construct transcripts they originate from, negating the need for prior knowledge of the sequences. Additionally, RNAseq analysis is facilitated by the availability of numerous sequenced genomes and their curated annotations.

Traditionally, RNAseq analysis involves mapping of short read sequences to a reference transcriptome or genome. Exon-intron boundaries are precisely inferred from mapped read densities, allowing the identification of canonically spliced transcripts. For read mapping, aligners are generally grouped into: spliced aligners and unspliced aligners (see (Garber et al. 2011) for a review). Spliced aligners can detect reads spanning exon-intron boundaries and split these reads to adequately map to the genome. For unspliced aligners, reads not mapping contiguously are generally discarded or sub-optimally aligned, introducing mismatches or gaps. Until recently, many chimeric transcripts (including PTES) remained undetected, as many studies excluded such transcripts or were only interested in characterising the expression of protein coding transcripts. Enrichment methods were also biased towards polyadenylated transcripts, in the process, excluding many circRNAs (Jeck & Sharpless 2014; Salzman 2016). The clinical importance of chimeric transcripts as possible biomarkers for cancers (and other diseases), has prompted the development of tools to identify chimeric RNA molecules. Many of these tools focused on identifying fused genes with parental genes on different chromosomes (Carrara et al. 2013; Chen et al. 2012; Liu et al. 2013; Hoffmann et al. 2014). However, more recently, algorithms for identifying PTES events have been published (Wang et al. 2010; Salzman et al. 2012; Salzman et al. 2013; Westholm et al. 2014; Memczak et al. 2013) . These tools commonly have a two step process of: 1) identify putative PTES events and 2) apply filters to exclude low confidence reads describing PTES events.

Approaches to identifying putative PTES events can be grouped into three:

1. Sequence Read Fragmentation Approach: This approach relies on the knowledge that PTES transcripts are defined by reads that can be mapped to two different regions/exons in

inverted (head-to-tail) order. Such reads do not map contiguously to the reference and would be unmapped or align sup-optimally. Thus, unmapped reads are routinely collected and fragmented into short segments (anchors) before remapping to the reference. Anchor alignments are then used to infer putative PTES events. Some tools generate anchor reads from termini of sequence reads (Memczak et al., 2013). Alignment of both anchors is expected to encompass the PTES junction. Other tools, such as MapSplice (Wang et al., 2010), generate multiple fragments from one read, attempt alignments for each fragment and define junctional sequence fragments as those without an alignment. These junctional fragments are then aligned locally after considering the alignment positions of neighboring fragments. The shortness of anchor reads presents a challenge, as short reads can align to multiple positions in the genome, posing a problem in discerning *bona fide* PTES events. This is can be considered a limitation of the fragmentation approach.

2. Paired-end Sequence Read Approach: In RNAseq, one end of cDNA templates can be sequenced to generate single-end sequence libraries. Paired-end reads are generated when both ends of cDNA templates are sequenced. By independently mapping PE reads, PTES events can be inferred from sequenced read pairs in inverted order. Akin to the anchor alignments (above), these PE reads encompass the PTES junction and to an extent, address the short read problem, since full length reads are used instead of short anchor reads. The distance between PE reads can also be used as a filter, excluding PTES events defined by PE reads with inner distance larger than expected. However, unlike the anchor reads approach, sequence between PE reads is undefined. Basically, because anchors originate from a single read, the intervening sequence between the anchors is known and can be used to infer the PTES junction. For PE reads, intervening sequences between PE reads are unknown and cannot be used to infer the structure of PTES transcripts with high confidence. In essence, it is difficult to establish whether PE reads originate from the same RNA molecule, thus, a potential limitation (Memczak et al., 2013). This approach is however used in many recently published methods (Salzman et al., 2012 & 2013; Zhang et al., 2014).

3. Brute-Force Approach: Another approach to PTES identification involves generating pseudo-sequence references for all possible combinations of exon-exon junction in shuffled order. For instance, a 3-exon gene would yield 6 putative PTES junctions: 1-1, 2-1, 3-1, 2-2, 3-2 and 3-3. Sequence reads are then aligned to these references to identify putative PTES events. This approach relies on existing knowledge of the transcriptome under study, thus, negates the discovery of PTES from unannotated loci. If not combined with adequate filters, this approach

may result in high false discovery rates, as alignments to the pseudo-references can be 'forced', producing gapped or sub-optimal alignments.

Pre-RNAseq *in silico* PTES identification methods utilized the brute-force approach and screened public databases of expressed sequence tags (EST). Dixon et al., (2005), produced 100bp fragments for all possible exon-exon combinations, mapped all fragments to EST using MegaBLAST (Altschul et al. 1990), selecting only fragments with >95% similarity to an EST. Fragments meeting that criterion were remapped to the human genome using BLAT (Kent 2002), only selecting fragments with two reported alignments and suggestive of PTES. This approach led to the identification of 263 PTES from 178 human genes, 98 in mouse; 17, 12, 27, and 8 in rat, chicken, zebrafish and fruit fly respectively. Similarly, Shao et al., 2006, generated 28bp sequence fragments with single base pair overlaps, from known human mRNA sequences and stored in an associative array. Using a sliding 28bp window, they screened EST sequences, comparing sequence fragments to stored mRNA sequences, to determine ESTs with fragments in head-to-tail orientation relative to mRNA. Their approach led to the identification of 817 human PTES transcripts (Shao et al. 2006). Following these efforts, several methods for PTES identification have been published. Table 1.2 summarizes methods described prior to the start of my study.

Study	organism/tissue	Data source	PTES identified
Dixon et al., 2005	human, mouse, rat, chicken, zebrafish and fruitfly	EST	425
Shao et al., 2006	human	EST	817
Danan et al., 2011	archaea	RNase R digested RNAseq	897
Al-Balool et al., 2011	human pediatric tumor samples	PolyA+ RNAseq	205
Salzman et al., 2012	human leukocytes	Ribosome-depleted total RNAseq	1319
Memczak et al., 2013	human, mouse and worm	Ribosome-depleted total RNAseq	4577
Jeck et al., 2013	human fibroblasts	RNase R digested and undigested RNAseq	7771

Table 1.2. **Published *in silico* PTES identifications.** List of reported *in silico* PTES identifications prior to my study (mid-2013). Earlier methods identified PTES from expressed sequence tags (EST).

1.3.5 Known sources of artefacts that confound in silico PTES identification

The primary challenge in *in silico* PTES identification is to distinguish between *bona fide* PTES events and false positive predictions. Various known sources of artefacts that may confound identification are introduced below.

Template-switching: In RNA-Seq experiments, RNA is first reverse transcribed into cDNA and amplified. Template Switching (TS) describes the scenario where the reverse transcriptase (RT) polymerase jumps to another template during cDNA synthesis (Cocquet et al. 2006; Odelberg et al. 1995). Houseley & Tollervey, (2010), demonstrated that homologous sequences around hairpin structures may result in RT jumping and exclude sequence within the hairpin akin to splicing. TS has also been shown to occur between sense and antisense transcripts of the same gene (Houseley & Tollervey 2010). The resulting TS transcript is subsequently sequenced; sequence reads from such transcripts may easily be confused as evidence for non-canonical splicing (Al-Balool et al., 2011; Wu et al., 2013).

Self-Priming: An additional source of artefact during cDNA synthesis involves self priming at termini of RNA templates. Lu et al., 2014 describes the phenomenon where small RNA molecules with hairpin structures at their termini can self-prime, particularly from their 3' ends and ligate to newly synthesized cDNA (Fig 1.7). The close proximity of both termini appears to aid this phenomenon, producing chimeric cDNA templates that mimic PTES. When sequenced, reads from these chimeric cDNA templates contain both ends of the RNA transcript and can confound *in silico* PTES identification.

Figure 1.7. Self-Priming of small RNAs. During cDNA synthesis of small RNAs with hairpin ends can self-prime (blue arrow) from 3' and ligated to newly synthesized cDNA (red arrow). In second strand synthesis, the ligated fragments can be used as templates, resulting in chimeric products. Figure adapted from Lu et al., (2014).

Multi-Locus Transcripts: Multi-locus transcripts may result from heterotypic/splice leader trans-splicing or gene fusions. As described above, fused genes may arise from chromosomal rearrangements or transcription-induced chimera events (Akiva et al., 2006; Nacu et al., 2011). In chromosomal translocations, genes around breakpoints are disrupted and (in some cases) fused. Additionally, unlike prokaryotes, monocystronic transcription occurs in eukaryotic cells. However, weak or disrupted transcription termination signals may lead to transcription read-through, producing transcripts comprised of more than one gene (Akiva et al., 2006). Subsequent splicing removes intergenic regions and join exons from constituent genes; reads from such transcripts may confound PTES discovery if there is high sequence identity between exons of both genes. For instance, paralogous genes such as *TUBA1A* and *TUBA1B*, which have exons with high sequence identity and are in close proximity in the genome, can produce chimeric transcripts that mimic PTES due to read-through transcription (Hansen et al. 2015).

Tandem Exon Duplication: At least 10% of human genes have duplicated exons (Letunic et al. 2002). In transcriptome-wide screens for PTES, a canonical splice between two exons (exons 2 and 3 for instance) may be confused for a single exon back-splice of exon 2 or exon 3 if both exons have high sequence identity. Furthermore, because only junctional reads are taken as evidence of PTES, theoretically, a high sequence identity between terminal sequences of two exons (exons 4 and 6 for instance) may result in reads supporting canonical splice between exons 5 and 6 to be mis-identified as evidence of PTES structure exon 5 - exon 4.

Certain positions in the genome consist of duplicated sequences that have evolved as a result of duplications of chromosomal segments (Samonte & Eichler 2002). There are 51,599 segmental duplications in the human genome (Bailey et al. 2002). These duplications typically have sizes ranging from 1Kb to >200Kb, have multiple copies of repeat sequences with over 90% sequence identity and are scattered all over the genome (Bailey et al. 2002). The repetitive nature of these duplications can be a source of artefacts and confound PTES discovery.

1.3.6 PTESFinder: a computational tool for PTES identification

As a result of limited availability of computational tools for PTES identification from high-throughput RNAseq data in 2012, I developed PTESFinder during my MRes project, building on initial scripts provided by Dr. Mauro Santibanez-Koref (Newcastle University). This pipeline combines the sequence read fragmentation approach with the brute-force approach, and is equipped with filters designed to exclude reads emanating from all known sources of artifacts. Briefly, the PTESFinder pipeline is split into three phases (Fig. 1.8A): 1) A discovery phase to identify putative PTES events from raw RNAseq data and generate putative PTES junction models; 2) an evaluation phase to examine the accuracy of predicted models and 3) a filtering phase (Izuogu et al., 2016).

Discovery phase: PTESFinder utilizes the read fragmentation approach to identify reads that map to the transcriptome in inverted order. Short sequence fragments (called anchors) are first generated from termini of raw sequences reads. These anchors are mapped to the transcriptome using Bowtie v. 1, a short read aligner (Langmead et al. 2009), with tolerance for only a single mismatch. To specifically identify intragenic PTES structures and eliminate sense-antisense template switching artefacts, alignment files in SAM format (Li et al. 2009) are processed further, requiring that both anchors from the same read map to the same gene in the transcriptome under study and in the same orientation, but in inverted order with respect to their order in the sequencing read. Anchor alignment positions are also used to determine constituting exons and in defining putative PTES transcript models (Fig. 1.8B).

Evaluation Phase: For each putative PTES junction model, new reference sequences (constructs) are generated by concatenating the last 65bp of the donor exon and the first 65bp of the acceptor exon, with the full exon sequence used if an exon is smaller than 65bp. To accommodate various read lengths, the segment sequence length used for model generation is adjustable, but should not exceed the read length, to allow for junction spanning read

alignments and filtering (introduced below). All reads within the dataset under study are then evaluated by re-mapping to these PTES constructs. This serves three purposes: First, anchor alignments are extended to ensure that the sequence between anchors is consistent with the predicted model; second, as RNAseq read lengths are short, this enables reads containing putative PTES junctions within the terminal sequence anchors to be accurately mapped; and lastly, it facilitates direct comparison with read mapping scores obtained from genomic and transcriptomic alignments during filtering (see below). Optionally, evaluation can also be 'guided' by supplying a database of previously discovered PTES structures, bypassing the requirement for model creation from reads under analysis.

Filtering Phase: To eliminate potential false positives originating from the genome under investigation, all the original reads are mapped to both genomic and transcriptomic references. The number of edits required for alignment (NM field in SAM format [Li et al., 2009]), and the number of perfectly aligned base pairs, are used to remove reads which align as well or better to either of these reference sequences than to the PTES constructs. To reduce template switching artefacts, which have heterogeneous junction points within short regions of often imperfect sequence homology (Houseley & Tollervey, 2010), reads which do not align perfectly to the exon junctions which define PTES are also removed using junctional filters. First, a user adjustable minimum junction span (JSpan) parameter is applied to ensure that there are no mismatches or gaps within 'n' nucleotides either side of the junction position, where n is an even integer. Second, to eliminate reads with regions of low quality alignment, a user adjustable segment percent identity (PID) parameter (see Izuogu et al., 2016 for details) is also applied independently to the segments on either side of the PTES junction, such that for a read to be retained both must meet or exceed the specified PID when aligned to the PTES construct. These user adjustable filters rely on alignment summaries provided by the NM field, MD field and Cigar in the SAM files (Li et al., 2009). The output includes the coordinates of the exon end involved in the junctions, a description of the PTES and the number of reads supporting the structure. This is presented in BED format (Kent et al. 2002). A second file contains additional information, read counts of all canonical exon junctions from transcripts where a PTES structure has been identified, to facilitate comparison with PTES counts.

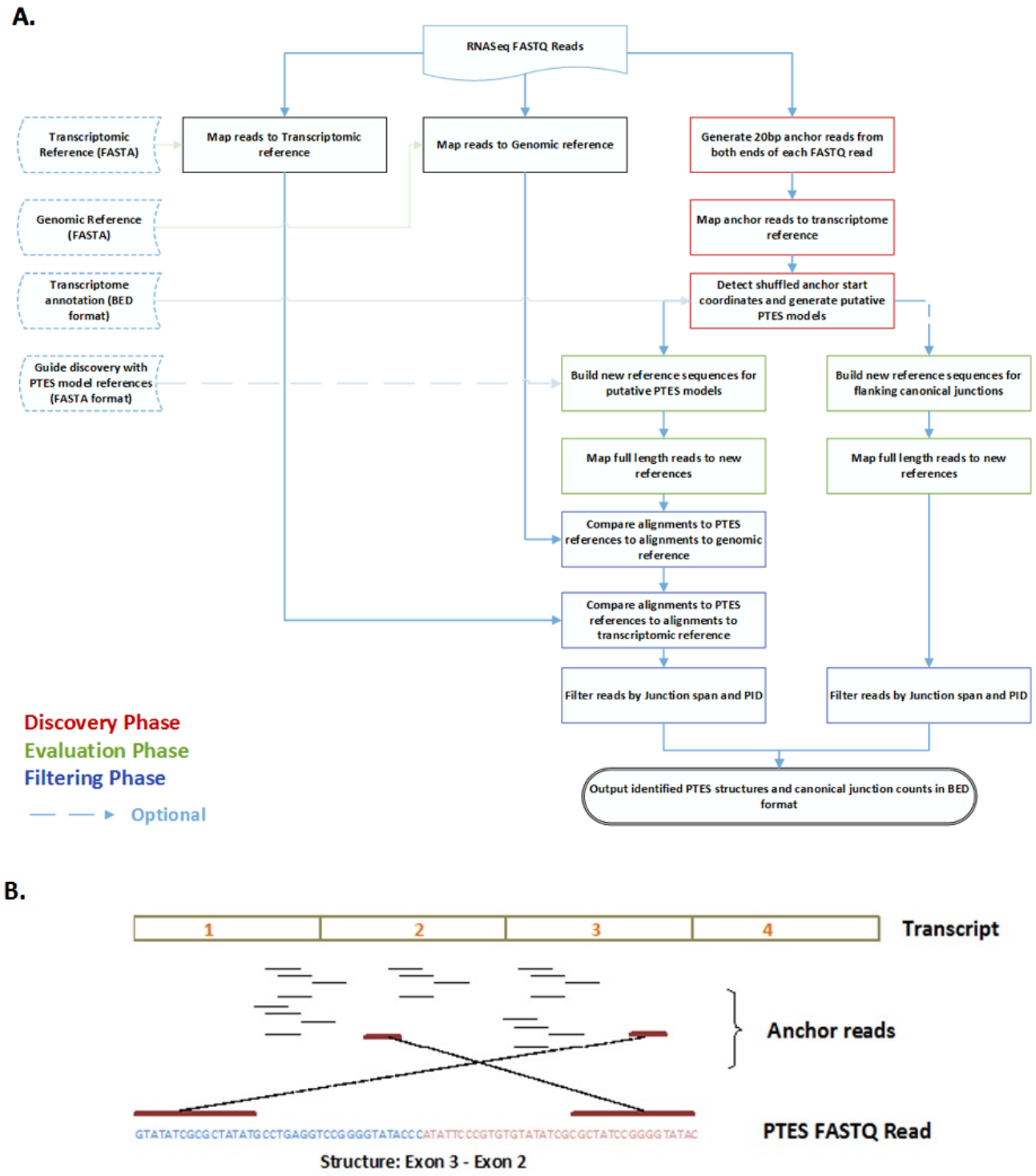


Figure 1.8. **Overview of PTES Discovery Pipeline.** A) The workflow includes three major phases: Discovery phase, Evaluation phase and Filtering phase. Putative PTES structures discovered using 20 bp anchor reads are evaluated by aligning full FASTQ reads to the models. The filtering phase includes stringent criteria designed to systematically exclude all known classes of false positive structures B) An illustration of split reads mapping used in the discovery phase. Figure and legend taken from Izuogu et al., (2016).

1.4 Project aims and outline of results chapters

Many questions pertaining to the precise mechanisms of PTES formation, their functional relevance and impact on mammalian transcriptomes remain poorly understood. How accurate are the proposed models for circularization? Are PTES transcripts exported via known nuclei-cytoplasmic pathways? Do these transcripts contribute to the proteome or have other functional significance? By what mechanism are they cleared from cells? How do they contribute to disease expressivity and development? This project aims to address these questions using both computational and experimental approaches.

Specifically, at the outset of this project, my aims were to first, assess the method for identifying PTES events from high-throughput RNAseq data, developed during my MRes. The first results chapter presents findings from assessment of various filters within PTESFinder, the effect of various aligner-specific parameters on PTES identification and reproducibility of predictions. Results of performance tests using both simulated data and RNAseq data, and comparisons with 4 published methods are also presented in this chapter. These have been published in a peer reviewed journal, BMC Bioinformatics (appendix 9.6).

Secondly, I use PTESFinder to analyze publicly available ENCODE RNAseq data from various cellular compartments, to define the distribution of PTES events in sub-cellular regions of 7 human cell lines. Using *in silico* methods, I also investigate the propensity for co-transcriptional PTES biogenesis and whether they are bound by polysomes, thus, translated.

In the third results chapter, I extend my investigation of PTES distribution in various cell lines to various human tissues and anucleate cells. Various properties of PTES transcripts in both platelets and mature erythrocytes are elucidated in this chapter using both *in vitro* and *in silico* methods. Work presented in this chapter resulted from extensive collaborations and have recently been published in the journal, Blood (appendix 9.6).

In the last results chapter, I investigate the temporal distribution of PTES transcripts, upon differentiation of human embryonic stem cells into retinal cells. Results from statistical analyses aimed at identifying PTES transcripts with functional roles in either pluripotency maintenance or differentiation are presented.

Chapter 2: Materials & Methods

In this chapter I describe *in vitro* and *in silico* methods used in analyses presented in this thesis. Any variations to methods and software parameters outlined in this chapter are explained within results chapters.

2.1 Cell lines

Table 2.1 describes RNAseq data sources (as defined by the ENCODE project) and human cell lines cultured. RNAseq data from mouse embryonic cell lines were also obtained and analyzed.

Cell line	Source description
DAOY	Derived from brain of male medulloblastoma patient
GM12878	Derived from B-lymphocytes of femal donor with normal karyotype
H1 & H9 ESC	Human embryonic stem cells
HEK293	Derived from embryonic kidney cells containing adenovirus 5 DNA
HeLa	Immortalised cell line from cervix of femal patient with cervical carcinoma
HepG2	Derived from liver of male patient with heptocellular carcinoma
Huvec	Derived from umblical vein endothelial cells with normal karyotype
K562	Imortalised cell line from blood of femal patient
NHEK	Derived from epidermal keratinocytes with normal karyotype

Table 2.1. **Description of Cell lines.** ENCODE project and ATCC (American Type Culture Collection) definition of human cell lines used in analyses presented in this thesis.

2.2 Sample Preparation

2.2.1 Tissue culture of HEK293 and DAOY cell lines

Cells were cultured in 75cm^2 tissue culture flasks at 37^{o}C in DMEM complete medium with 100 IU/ml penicillin, 100µg/ml streptomycin and 10% heat activated fetal bovine serum (FBS) from Sigma-Aldrich (Dorset, UK).

2.2.2 Differentiation of H9 ESC

Human embryonic stem cells (H9 line, passage 34-45), were expanded and differentiated by Prof. Lako (Newcastle University, UK) and colleagues, using their previously described

protocol published in Mellough et al., (2012), with slight modifications published in Mellough et al., (2015). Cells were differentiated in triplicate with and without insulin growth factor 1 (IGF-1) treatment.

2.2.3 Human tissues and blood samples from healthy donors

RNA from various human tissues were obtained from Biochain (Amsbio, UK). Red blood cells (RBC) were obtained from the Scottish National Blood Transfusion Service in accordance with the terms of the standard donor consent form and information and with approval of the Scottish National Blood Transfusion Service Sample Governance Committee. Cord blood was obtained by Dr. Cedric Ghaveart and colleagues (Cambridge University, UK) after informed consent under a protocol approved by the National Research Ethics Service (Cambridgeshire 4 Research Ethics Committee ref. no. 07/MRES/44). Whole blood samples were collected from healthy volunteers with approval from Newcastle University's Faculty of Medical Sciences Ethics Committee (909/2015). Platelets, platelet-rich plasma (PRP), peripheral blood mononuclear cells (PBMCs) and RBCs were isolated by Dr. Alhasan (see Alhasan et al., 2016 for details).

2.2.4 RNA Isolation and cDNA synthesis

RNA was extracted using Trizol (Life Technologies) and treated with DNase I (Promega, Southampton, UK) according to manufacturer's instructions. Briefly, cells were washed with ice cold PBS, before adding 1ml Trizol and left at room temperature (RT) for 5 mins. Cells were then transferred into a 1.5ml eppendorf tube using a cell scraper and left for further 5 mins at RT. Afterwards, 0.2ml of chloroform (in 1:5 ratio to Trizol volume) was added, shaking vigorously for 15 secs, before leaving on ice for 3 mins and centrifuging for 15 mins at 1200g in 4^{o}C. The upper aqueous layer was then transferred to an eppendorf tube and mixed with equal volume of isopropanol, left on ice for 10 mins and centrifuged at 1200g for 10mins. The supernatant was then discarded and pellets washed with 1ml of 75% ethanol (equal volume of Trizol) and further centrifuged for 5 mins at 7500g. Following ethanol precipitation, the supernatant was removed, leaving the pellet to air dry, before re-suspending in 50µl of Diethylpyrocarbonate (DEPC) treated water.

For cellular fractionation and RNA extraction from sub-cellular compartments, Cytoplasmic and Nuclear RNA extraction kit from Norgen Biotek (https://norgenbiotek.com) was used, instead of the Trizol extraction protocol described above. RNA quantification was performed using NanoDrop ND-1000 spectrophotometer (ThermoFisher Scientific - http://www.thermofisher.com) and quality was assessed using an Agilent 2100 Bioanalyser

(Agilent Technologies, California, US). cDNA was synthesised using high-capacity cDNA generation kits (Applied Biosystems - http://www.thermofisher.com), random hexamers and Moloney murine leukaemia virus (MMLV). For comparisons with AMV reverse transcriptase generated cDNA, an AMV RT kit from Promega were obtained and used in cDNA synthesis.

2.2.5 RNase R digestion

One microgram of RNA was added to 1µl of 10X RNase R buffer in DEPC H_2O and 20 units of *E. coli* RNase R (Epicentre Biotechnologies - www.epibio.com), or zero units (1µl of DEPC H_2O) for mock treatment in a 10µl reaction volume. Tubes were then incubated at 37^{o}C for 30 minutes. Ethanol precipitation and cDNA generation were performed as outlined above.

2.3 *In vitro* PTES confirmation, visualization and quantification

2.3.1 Primer design

Oligonucleotide primer pairs in inverted orientation and spanning PTES junctions were designed using the primer-blast tool from NCBI (Ye et al., 2012). For each PTES transcript, exonic sequences around the junction were concatenated and used as input, requiring a minimum amplicon size of 150bp in most cases and using default values for optimal melting temperature. The specificity of generated primer pairs was checked by *in silico* PCR (Kent 2002). No amplicons of the expected size should be observed from the genome or transcriptome. Primer pairs meeting this specificity requirement were purchased from Metabion International AG, Germany (see appendix 9.1 for all primer sequences).

2.3.2 Polymerase Chain Reaction (PCR)

Master Mix reagents: For each sample, 19µl master mixes were prepared with reagents from Promega (Madison, USA):

- 4µl 5X GoTaq Green Buffer
- 2µl dNTP mix (0.2mM per dNTP)
- 1.5µl 10pM Forward primer
- 1.5µl 10pM Reverse primer
- 9.7µl sterile H_20
- 0.3µl *Taq* Polymerase
- 1µl of cDNA template

PCR Cycle (using Sensoquest thermal cycler from GeneFlow, Staffordshire, UK):

- Initial denaturing at 95^{o}C for 2 minutes
- Denaturing at 95^{o}C for 1 minute
- Primer annealing at 56^{o}C for 1 minute
- DNA synthesis at 72^{o}C for 1 minute; return to denaturing to repeat for 35 cycles
- Final DNA synthesis at 72^{o}C for 1 minute
- Store at 4^{o}C

2.3.3 Agarose Gel electropherosis

2% agarose gels comprising of: 100mL of TBE buffer and 2g of agarose were prepared and stained with GelRed^{TM} (Biotium, Hayward, USA) nucleic acid gel stain. Gels were prepared in 15 x 10 cm gel trays and ran in sub-cell GT cell gel tanks (BioRad, Hemel Hempstead, UK). Run parameters: 95 volts for 50 minutes and visualized under UV light.

2.3.4 Quantitative PCR (qPCR)

Quantitative PCR experiments were performed by Dr. Alhasan (Newcastle University, UK) using Taqman master mix (Life Technologies). Transcript expression was normalized using the ΔCT method relative to the geometric mean of 4 housekeeping genes (*GAPDH*, *PPIA*, *TUBB*, *GUSB*) analyzed using TaqMan gene expression assays (Applied Biosystems - http://www.thermofisher.com)(see Alhasan et al., 2016 for details of assays). Circular transcripts were also normalized against the linear transcript from the same gene where appropriate. Reactions were performed in 10 µl volumes in 384 plates using QuantStudio 7 Flex (Applied Biosystems) with the following cycling parameters: 2 minutes at 50^{o}C, 10 minutes at 95 ^{o}C, followed by 40 cycles of 15 seconds at 95 ^{o}C and 60 ^{o}C for 1 minute.

2.4 Public RNAseq datasets

2.4.1 Human Fibroblasts and Leukocytes data

Publicly available RNAseq data generated from Leukocytes (n = 6 [Salzman et al., 2012]), HEK293 (Memczak et al., 2013) and Fibroblasts (n = 4 [Jeck et al., 2013]) were obtained from an NCBI sequence read archive mirror: http://sra.dnanexus.com. Leukocytes (CD19+, CD34+ & neutrophils) and HEK293 samples were generated from ribosome depleted

total RNA extracts, sequenced on Illumina HiSeq 2000 platform and deposited under the following SRA ids: SRR364679-81 and SRR384963-5. Fibroblast samples were either digested with RNase R (SRR444974 & SRR4445016) or undigested (SRR444975 & SRR444655). With the exception of HEK293 and Fibroblast samples, which had 100bp reads from paired-end libraries, all samples had 76bp reads from single-end libraries.

2.4.2 ENCODE sub-cellular RNASeq data

RNAseq data published by the ENCODE consortium and available through NCBI Gene Expression Omnibus (GEO - http://www.ncbi.nlm.nih.gov/geo/), under the data accession: GSE30567, were obtained. Twenty-three samples were generated from non-polyadenylated (PolyA-) long RNA (>200bp) transcripts extracted from the nuclei (n = 13) and cytosol (n = 10) of various human cell lines (see appendix 9.1). Four samples from the nucleus and cytosolic RNA PolyA+ fractions of GM12878 and K562 cells were also obtained. Additionally, 2 samples each from the nucleoplasm, nucleoli and chromatin of K562 were also obtained and analyzed. All samples were paired-end libraries with 76 bp reads.

2.4.3 Sucrose-gradient fractionated RNAseq data from HEK293

RNAseq data from Karginov and Hannon (2013) (GEO accession: GSE44404), generated after sucrose gradient fractionation of HEK293 cells, were obtained from an NCBI sequence read archive mirror: http://sra.dnanexus.com. In total, 16 single end short read (50bp) samples from that study were analyzed, eight were treated with arsenite to induce translational arrest (SRA ids: SRR742818 - 25) and 8 were control samples (SRA ids: SRR742810 - 17).

2.4.4 RNAseq data from human tissues and anucleate cells

Ribosome depleted RNAseq data from anucleate cells, nucleated tissues, human cell lines and samples from cell lines treated with RNase R were obtained from http://sra.dnanexus.com. Additionally, 2 polyA+ samples from platelets and megakaryocytes were obtained. With the exception of the polyA+ samples, all samples were paired-end. Table 2.2 shows sample ids and sources.

Sample Id	Sample	RNA Fraction	Reference
SRR444974	Fibroblasts_D1	RNaseR Digested	(Jeck et al., 2013)
SRR445016	Fibroblasts_D2		
SRR901967	H9_digested		(Zhang et al., 2014)
SRR444975	Fibroblasts_U1	Total (ribosomal RNA depleted)	(Jeck et al., 2013)
SRR444655	Fibroblasts_U2		
SRR768411	GM12878_1		(Birney et al., 2007)
SRR768412	GM12878_2		
SRR768413	K562_1		
SRR768414	K562_2		
ERR335312	Platelets_F		(Kissopoulou, Jonasson, Lindahl, & Osman, 2013)
ERR335311	Platelets_M1		
ERR335313	Platelets_M2		
SRR2038798	RBC		Institute of Genetic Medicine, Newcastle, UK
SRR787270	Bladder		(Nielsen et al., 2014)
SRR787271	Brain		
SRR787272	Breast		
SRR787273	Colon		
SRR787274	Heart		
SRR787275	Kidney		
SRR787276	Liver		
SRR787277	Lung		
SRR787278	Muscle		
SRR787279	Ovary		
SRR787280	Prostate		
SRR787281	Skin		
ERR065725	Megakaryocytes	Poly A+	(Nürnberg et al., 2012)
ERR039697	Platelets		(Kissopoulou et al., 2013)

Table 2.2. **RNAseq data from anucleate cells and nucleated tissues**. SRA ids of publicly available RNAseq data from anucleate cells and human nucleated tissues.

2.5 RNAseq data generation

2.5.1 High-throughput RNA sequencing

RNA from mature erythrocytes and three time points (days 0, 45 & 90) following differentiation of human embryonic stem cells (hESC) were ribosome depleted and sequenced by AROS Applied Biotechnology (Aarhus, Denmark), using the TruSeq RiboZero Stranded mRNA LT kit (Illumina - http://www.illumina.com). RNA quantification results for all sequenced samples are presented in Table 2.3 below. Sequencing produced paired-end 100bp sequence libraries.

Sample Information			Bioanalyzer Analysis		Nanodrop Analysis	
Sample	Tissue / Cell line	RIN	Concentration (ng/µl)	A260	A280	260/280
Day 0	H9 Embyonic Stem Cells	9.80	1499.10	37.48	18.09	2.07
Day 45 Control		8.10	50.20	1.25	0.63	2.00
Day 45 IGF-1		9.10	1453.70	36.34	17.39	2.09
Day 90 Control		7.90	87.80	2.19	1.08	2.03
Day 90 IGF-1		8.20	33.40	0.83	0.42	1.97
Day 0		10.00	2387.80	59.69	28.78	2.07
Day 45 Control		8.60	76.90	1.92	0.94	2.04
Day 45 IGF-1		8.40	46.20	1.16	0.57	2.03
Day 90 Control		7.10	134.90	3.37	1.65	2.04
Day 90 IGF-1		7.10	35.10	0.88	0.42	2.08
Day 0		10.00	1731.70	43.29	20.84	2.08
Day 45 Control		7.60	102.00	2.55	1.23	2.07
Day 45 IGF-1		8.30	140.50	3.51	1.70	2.07
Day 90 Control		8.90	30.60	0.76	0.39	1.96
Day 90 IGF-1		8.30	18.60	0.46	0.23	2.04
RBC-RNA-1	RBC	6.80	57.10	1.42	0.67	2.14

Table 2.3. **RNA extracts from mature erythrocytes and H9 ESC differentiation series**. Quality metrics of RNA extracts from H9 embryonic stem cells and Red Blood Cells (RBC), prior to sequencing.

2.5.2 Generating simulated RNAseq data

All published PTES structures within circbase.org (Glažar et al. 2014) and in common with structures identified from human fibroblasts (Jeck et al., 2013) using PTESFinder were obtained. In 100 simulations, ~5000 PTES structures were randomly selected along with 5000 canonical junctions and pooled. Synthetic reads were generated from these structures in each simulation at different depths of coverage. To generate these reads, constructs were generated for each junction, by concatenating the full sequence of both exons, and 100bp reads were then generated from each sequence at read depths of 2, 5, 10, 25 and 50. Even coverage was achieved by segmenting each construct by read length and randomly choosing read start positions within each segment. For instance, for a desired coverage of 5 and a construct of 430bp, 5 random start positions are chosen from each of the pre-defined ranges (0 -100bp, 100 -200bp, 200 – 300bp & 300 – 430bp) (illustrated in Fig. 2.1). Reads spanning junctions after runs were recorded for subsequent comparisons.

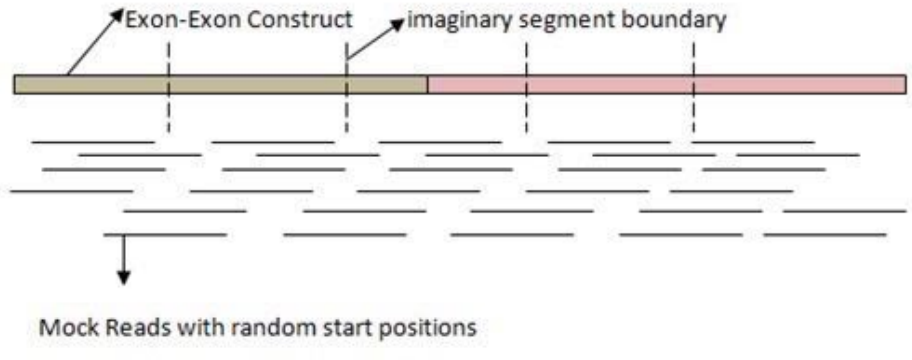


Figure 2.1. **Methodology for simulating reads.** 100bp reads were simulated from constructs, with random read start positions to ensure even coverage across constructs

2.5.3 Sub-sampling of RNAseq data

Sub-sampling of reads, at 25% and 75% of library size, was performed using SEQTK (https://github.com/lh3/seqtk), using the following command: *seqtk sample -s$RANDOM sample.fastq 0.25 > subsample.fastq* .

2.6 Computational Methods

2.6.1 Sequence Quality check

Quality of sequenced reads was checked with FASTQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/). Where necessary, reads were trimmed to remove segments with poor sequence quality prior to downstream analysis.

2.6.2 PTES identification

Paired-end reads were merged after modifying read ids. PTESFinder v.1 was used to screen all sequence data for PTES transcripts with the following parameters: JSpan = 8, PID = 0.85, segment size = 65 (25 for short reads) and m = 7. In most cases, analyses were guided by supplying FASTA sequences of previously identified PTES junctions (n = 40594). Guided analysis is necessary when identification with anchor reads is not meaningful, as is the case for very short reads. Bowtie (Langmead et al., 2009) indexes for HG19 were used for alignments to the genome; RefSeq (catalog 58 & 65) and GENCODE v. 19 transcript models were obtained from UCSC genome table browser (Kent et al., 2002) on 20/04/2013 and 14/07/2014 respectively.

To assess PTES predictions unconstrained by curated splice junctions, PTESFinder was modified. Two versions were developed to investigate predictions using both spliced and unspliced aligners. In both versions, the anchor mapping phase was modified, replacing anchor generation and bowtie1 alignment with STAR v. 2.4 (Dobin et al. 2013) soft-clipped alignment.

In one version, Bowtie2 (Langmead & Salzberg 2012) was replaced as the choice of aligner for alignment of full length reads to the genome.

Additionally, scripts published by Memczak *et al.*, 2013 (default parameter values), CIRI v. 1.2 (default parameter values [Gao et al., 2015]), circRNA_finder (default parameter values [Westholm et al., 2014]), and MapSplice v. 2.1.5 (Wang et al., 2010) used in Jeck et al., (2013) (parameters: --fusion --non-canonical -p16), were obtained and used in screening RNAseq data for PTES.

2.6.3 RNAseq analysis

Sequence references: Genome and transcriptome FASTA files for human (HG19) and mouse (MM10) were obtained from UCSC genome browser (Kent et al., 2002). Aligner-specific index files were built for STAR (command: STAR —runThreadN 8 —runMode genomeGenerate —genomeFastaFiles $hg19_dir —genomeDir $output_dir), Bowtie1 (command: bowtie-build hg19.fa hg19) and Bowtie2 (bowtie2-build hg19.fa hg19) aligners. Sequence references for all exon-exon, exon-intron and terminal exons were also generated and indexes built using bowtie2.

Mapping: Sequence reads were first mapped to the genome using Bowtie2 (Langmead & Salzberg, 2012) to derive inner distance metrics prior to Tophat (Trapnell et al. 2009) runs. Metrics were calculated using *CollectInsertSizeMetrics.jar* from Picard (https://sourceforge.net/projects/picard/). In addition to derived inner distance metrics, parameters for Tophat runs include: —library-type fr-firststrand, —no-coverage-search, —b2-sensitive, —microexon-search and -x 20. For alignments to the genome using STAR (Dobin et al., 2012), parameters used were: —outFilterMultimapNmax 7 and —outFilterMismatchNmax 2.

Expression estimates: HTSeq (Anders et al. 2014) and Cufflinks (Trapnell et al. 2012) were used to quantify transcripts. Exon RPKMs were derived by extracting reads mapped within genomic coordinates of exons, using CoverageBed from BEDTools (Quinlan & Hall 2010). Counts (C) were normalized by size (S) and total mapped reads (M), using this formula: (C / S) * (10E9 / M). In one analysis, reads were mapped to ERCC (a standard set of exogenous synthetic RNAs used as controls in gene expression analysis (Jiang et al. 2011; ERCC 2005)) spike-in sequence references and library size scale factors were estimated using *estimateSizeFactors* from DESeq2 (Love et al. 2014). Scale factors were used to normalize PTES raw counts and terminal exon sequences.

Cluster analysis: Expression estimates of PTES and canonical junctions were used to assess similarities between samples. Hierarchical clustering of samples was performed using

Euclidean distance between samples, derived using expression estimates. KMeans clustering was used to assess expression changes between samples and distinct expression profiles of linear and circRNAs. To determine the number of clusters, K, the elbow method was used. For this method, initial cluster analyses using a range of K values, from 2 to 15 were performed. The final choice of K was determined by the lowest value of K that maximizes the proportion of variance explained by clustering, with increasing K having little additional effect.

Differential expression analysis: Statistical tests of differential expression (based on raw counts produced using HTSeq) were performed using DESeq2 (Love et al., 2014) package in R (https://www.r-project.org), a free software tool for statistical computing.

Visualization: BigWig files were generated from alignments to the genome using *genomeCoverageBed* from BEDtools and *bedGraphToBigWig* from UCSC. BigWigs were then archived and served through Galaxy, a web-based tool for sequence analysis (Afgan et al. 2016). Distributions of aligned reads were visually examined on the integrative genomics viewer (IGV v 2.1.21 (Robinson et al. 2011)) and on the UCSC genome browser.

Genomic features overlap: All known RNA editing sites within rnaedit.com database (Porath et al. 2014; Chen 2013), Bisulphite data (in bedGraph format) from H9 differentiation into retinal pigment epithelium (Liu et al. 2014) and miRNA binding sites (miRcode - (Jeggari et al. 2012)) data were obtained on 5/02/2016, 12/10/2015 and 12/05/2015 respectively. Overlap and proximity of genomic features were accessed by comparing their genomic coordinates using *intersectBed* and *closestBed* from BEDTools respectively. Genomic coordinates specific to HG18 build were converted to HG19 coordinates using the liftOver tool from UCSC genome browser (Kent 2002), to allow for direct comparisons.

2.6.4 Definition and derivation of metrics

$RPKM_I$ and $RPKM_E$: For each gene, exons predicted to lie within any circRNAs identified from that gene (in any sample) were used to estimate $RPKM_I$; exons external to all circRNAs in all samples were used to derive $RPKM_E$. In both cases, read counts of exons were summed and normalised by total size of exons. RPKMs were derived using this formula: (C / S) * (10E9 / M); where C is the total read counts, S is the total size and M is the total mapped reads in respective samples. $RPKM_I$ is an estimate of expression of both linear and circular RNAs from each locus and $RPKM_E$ is an estimate of expression for linear molecules only.

Abundance ratios (AR): This is a measure of the abundance of identified PTES junctions relative to canonical junctions. This metric is derived in three ways: 1) dividing PTES junction

counts by total canonical junction counts observed from host locus; 2) computing the ratio: $RPKM_I$ / $RPKM_E$ or 3) $RPKM_E$ / ($RPKM_E$ + $RPKM_I$).

Co-transcriptional splicing rates (CSR): To estimate the rate of co-transcriptional splicing, expression estimates of chromatin-associated pre-mRNAs and analogous mRNAs were compared. For mature mRNAs, sequence references for all exon-exon junctions between the first exon and other exons, and between other exons and last exon of the longest isoforms of each gene with at least 3 exons were generated. For pre-mRNAs, sequence references of exon-intron junctions for first and last exons were also generated. Reads from K562 chromatin associated samples were aligned to these junctional references. Reads supporting all splice junctions involving first and last exons of genes were extracted, accepting only mapped reads with edit distance <= 2. Read counts supporting exon-exon junctions were extracted and aggregated. Reads mapped to exon-intron junctions involving the terminal exons were also extracted. The co-transcriptional splicing rate (CSR) for each gene was derived using the formula: CSR = X / (X + Y); where X = sum of read counts from canonical splice junctions involving first and last exons, and Y = reads counts from exon-intron junctions involving first and last exons.

Percentile read coverage: Even distribution of reads across each transcript (linear and circular) was examined by first generating 100 sequence segments of sizes relative to transcript length. Uniform read coverage across linear transcripts was assessed using RSeQC v. 2.4 (Wang et al. 2012). For circRNAs, transcript lengths were determined by concatenating all exons predicted within each circRNA. The number of nucleotides covered by at least one read within each of the 100 segments was used to derive the read coverage for that sequence segment. For instance, each sequence segment of an 1000bp transcript will have length of 10bp. The read coverage of the first segment of this hypothetical transcript will be 80%, if only 8bp of the 10bp are covered by at least one read within the RNAseq data. Comparisons between samples using this metric were performed by computing the difference in coverage for each segment per transcript.

MiRNA binding sites density: Genomic coordinates of exons were compared with genomic positions of miRNA binding sites, counting number of binding sites within each exon. For each PTES transcript, exons predicted to be within the transcript are used to determine the number of miRNA binding sites within the transcript. The spliced size in bp (derived by concatenating exons within circRNA) was used to normalize derived counts, resulting in miRNA binding site density.

Software performance: For simulated data, transcripts correctly identified by each PTES identification method after each simulation, were determined by comparing genomic coordinates of identified transcripts with the genomic coordinates of transcripts expected to be

recovered from within each dataset. The numbers of correctly identified PTES transcripts (true positives – TP), incorrectly identified PTES transcripts (false positives – FP), PTES transcripts incorrectly excluded (false negatives – FN), and canonical junctions correctly excluded (true negatives – TN), were used to estimate **sensitivity:** TP / (TP + FN), **specificity:** TN / (TN + FP), and **false discovery rate** (FDR): FP / (TP + FP). All analyses were carried out on a high performance cluster consisting of 20 nodes, each having 8 CPU cores, running at 2.67 GHz. Sixteen of these nodes have 48 GB memory, while the other 4 have 96 GB memory. Run times and memory consumption of each method were recorded for comparisons.

2.6.5 Statistical analysis of PTES abundance

Reads counts for PTES transcripts identified from biological replicates were summed to reduce the effect of sampling on downstream analyses. Total canonical junction counts from PTES producing genes were also tallied. Two-by-two contingency tables - consisting of PTES counts and canonical junctions counts for samples being compared - were derived and used in Fisher's exact tests for each transcript. Fisher's exact test was chosen because of its tolerance for low values in calculating the significance of deviation from the null hypothesis; null hypothesis being no difference in PTES distribution is expected between sample groups. Multiple testing correction using Benjamini-Hochberg (BH) method was then performed. For t-tests, PTES raw counts from each sample were first normalized by dividing with total canonical junction counts from their respective host locus or by total junction counts (PTES and canonical) from respective samples.

Enrichment analysis using $RPKM_I$ / $RPKM_E$ ratios were performed using the Wilcoxon signed-rank test after removing genes with less than 4.0 RPKM for PTES exons. False discovery rate was controlled at 0.01 using Benjamini-Hochberg method. All statistical analyses were performed using R statistical computing software versions 2.15.1 and 3.1.1.

2.6.6 Custom scripts

Custom scripts were developed for extracting read counts, genomic features, enrichment analyses and derivation of all metrics used in downstream analyses presented in this thesis. Scripts were developed using Java 2 standard edition (J2SE), R statistical software and shell scripting. A list of all scripts developed, their expected inputs and outputs is presented in appendix 9.1.

Chapter 3. Assessment of Computational PTES identification Methods

3.1 Introduction

The first PTES transcripts to be described were identified fortuitously using *in vitro* methods (Nigro et al., 1991; Cocquerelle et al. 1992; Cocquerelle et al. 1993; Bailleul 1996). The emergence of high throughput RNA sequencing (RNAseq) allows for the simultaneous identification and quantification of various RNA species, presenting the opportunity to characterise PTES transcriptome-wide. Many computational tools for PTES identification from RNAseq data have now been described (Memczak et al., 2013; Salzman et al., 2012 & 2013; Jeck et al., 2013), but most do not directly address the primary challenge of discriminating between *bona fide* PTES predictions and artefacts.

3.1.1 Existing PTES identification tools do not specifically exclude all sources of artefacts

A recent report found that up to 50% of previously reported PTES transcripts are artefacts (Yu et al., 2014). Sources of false positive reads include: reads emanating from pseudogenes, segmental duplications, tandem exon duplications, fused genes; read-through transcripts and template-switching during cDNA synthesis (see 1.3.5 for details). A common approach for eliminating false positive reads originating from other genomic regions is an initial mapping to the genome and screening only unmapped reads for PTES events, thus reducing the likelihood of misidentification. This approach is utilized by many existing tools for PTES identification (Memczak et al., 2013; Zhang et al., 2014; Wang et al., 2010; Guo et al., 2014; Westholm et al., 2014). However, many classes of false positive structures can elude this filter. For instance, Memczak et al., 2013 reported the identification of 1950 PTES transcripts from human leukocytes and HEK293 cells. Reads supporting 7 of the 20 most abundant circRNAs identified in that study map with high confidence to linear RefSeq (Pruitt et al., 2002) entries. Reads supporting canonical splice events can be mischaracterized as PTES supporting reads due to high sequence identity between exons, as exemplified by 4 circRNAs reported in that study (Fig 3.1, Appendix 9.2). CircRNAs with high sequence identity with spliced pseudogenes have also been reported. These spliced pseudogenes could conceivably arise from retrotransposition of chimeric RNAs, transposition from within segmental duplications or tandem exon duplication events.

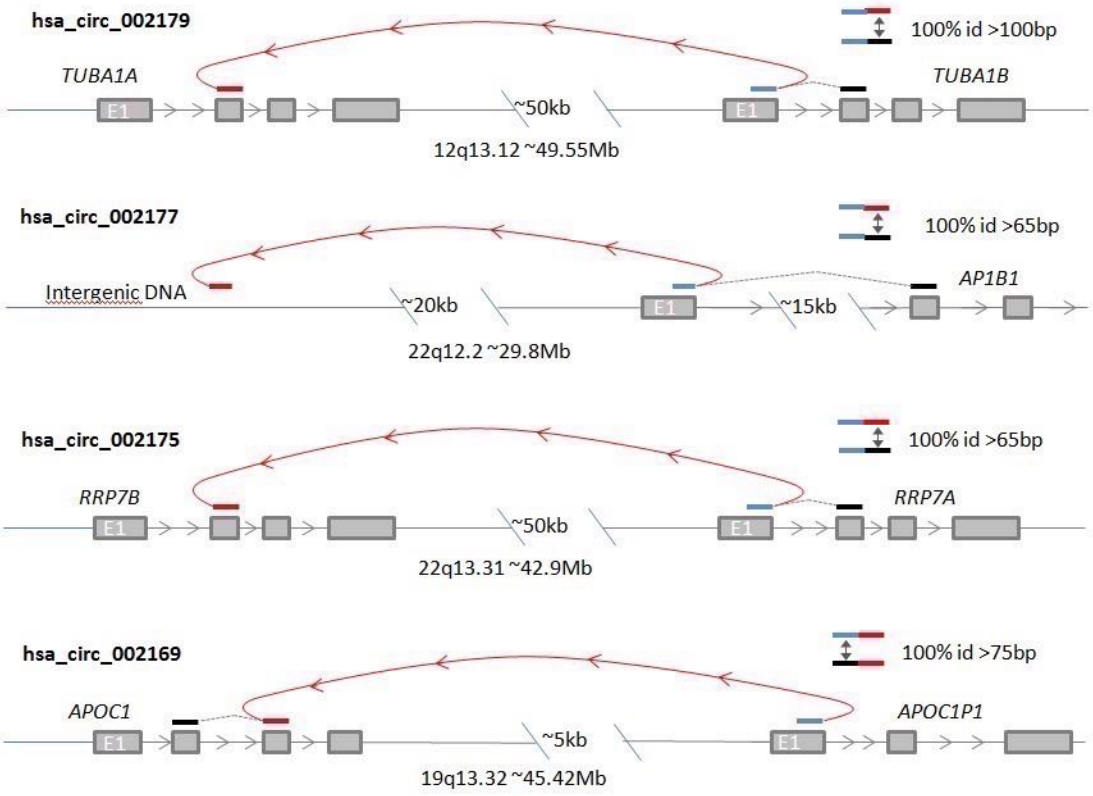


Figure 3.1. **Examples of Intragenic False Positives**. Schematic diagrams showing inferred structure and key sequence relationships for 4 of the 20 most abundant circRNAs reported in Memczak et al., (2013). In each case, the inferred structure shares 100 % identity to a linear transcript spanning the defining exon-exon junction. Within the top 20, hsa_circ_002174, 002165 and 002164 show similar patterns of identity to multiple genomic locations. Blue – Inferred Donor Exon, Red – Inferred Acceptor Exon, Black – upstream or downstream RefSeq exon sharing 100 % identity to donor or acceptor exon at junction. Approximate chromosomal locations (HG19) are shown. Figure and legend taken from Izuogu et al., (2016).

3.1.2 Choice of aligner and aligner-specific parameters may impact reproducibility of PTES predictions

Various reports have highlighted apparent low overlap between the PTES transcripts identified from the same samples using different methods (Yu et al., 2014; Hansen et al., 2015; Chen et al., 2015). Low overlap in predictions is suggestive of low specificity in existing methods (Chen et al. 2015). Virtually all published computational methods for PTES identification relies on established aligners for initial mapping of reads to the genome or transcriptome. Aligners can be splice-aware (e.g. Tophat (Trapnell et al. 2009), STAR (Dobin et al., 2013) etc.), recognizing reads that span splice junctions, subsequently splitting such reads to produce accurate alignments. Other aligners (like Bowtie (Langmead et al., 2009), BWA (Li & Durbin 2009) etc) are not splice-aware, and instead are best suited for mapping short reads contiguously, introducing gaps and mismatches for most reads that span splice junctions. It

follows that the choice of aligner can directly affect the performance of PTES identification tools.

Furthermore, within each aligner, certain parameters are used to determine the suitability of reported alignments. Typically, a read with alignments to many targets can either have no reported alignment, a randomly selected alignment reported or multiple reported alignments. Where reported, such alignments may be used to wrongly support a putative PTES event, as the true origin is ambiguous. Existing tools arbitrarily assign values to aligner-specific parameters that guide whether alignments to multiple targets are reported. For instance, Salzman et al., (2012) allowed alignments to multiple targets and up to 3 mismatches in alignments. In contrast, Memczak et al., (2013) accepted alignments to fewer than 20 targets and a maximum of 2 mismatches in alignments. Both methods differ in the number of PTES transcripts identified from the same leukocytes RNAseq data, presumably as a result of these aligner-specific parameters and respective filtering criteria.

3.1.3 PTESFinder is equipped with filters that systematically exclude sources of artefacts

To fully characterize PTES transcripts, define their global properties and functional relevance, an accurate identification method is required. To that end, I developed a method, PTESFinder during my MRes. This method screens all reads, not just those that fail to map to the genome and is equipped with filters designed to systematically exclude all known sources of artefacts. First, to identify putative PTES events, only anchor reads aligned in reverse orientation to the same locus are accepted as initial evidence for PTES. This stipulation reduces the likelihood of mischaracterizing reads from other chimeric transcripts as evidence for PTES. Subsequently, 3 main filtering criteria are applied, targeting other known sources of artefacts. Alignment qualities of reads mapped to predicted PTES junction models (constructs) are compared with alignment qualities of same reads when mapped to the genome - genomic filter - and when mapped to the transcriptome - transcriptomic filter (see 1.3.6 for details). Both filters were designed to target false positive reads emanating from pseudogenes, segmental duplications, tandem-exon duplications and other genomic features. An additional filter, junctional filter, is then applied to target false positive reads from template-switching events. The qualities of alignments around the non-canonical junctions which define each PTES are evaluated using minimum junctions span (JSpan) and segment percent identity (PID). The JSpan parameter is a required even integer value that is used to eliminate reads with sub-optimal alignments around the PTES junction. For instance, no mismatch or gap is allowed 3bp either

side of PTES junctions when a JSpan of 6 is specified. The PID parameter is a required float values that is used to eliminate reads with low sequence identity to PTES construct. A PID of 85% will exclude reads with less than 85% sequence identity to sequence either side of the putative PTES junctions.

Unlike many existing PTES identification tools that do not require curated annotations of the transcriptome under study, PTESFinder is designed to restrict PTES detection to backsplices occurring at known splice junctions. This restriction excludes PTES transcripts originating from intergenic and intronic regions, and backsplices utilising cryptic splice sites or non-canonical splice signals.

3.2 Aims

In addition to potential false positive PTES predictions by existing tools, there is, to varying degrees, a lack of concordance in predictions made by various methods, further highlighting the need for an accurate computational method and an assessment of diverse approaches. In this chapter, my specific aims are:

- Assess the effectiveness of filters within PTESFinder.
- Compare PTESFinder to published PTES identification tools
- Investigate aligner types and aligner-specific parameters as a source of variation between different tools

3.3 Results

As an initial assessment of PTESFinder function, RNAseq data from human fibroblasts total RNA which has previously been mined for circRNAs (sample SRR44975A in Jeck et al., (2013)), were analyzed both with and without the application of the genomic and transcriptomic alignment filters. Reads excluded during analyses, together with alignment edit distances of reads identified by each filter applied separately, are shown in Figure 3.2A-B.

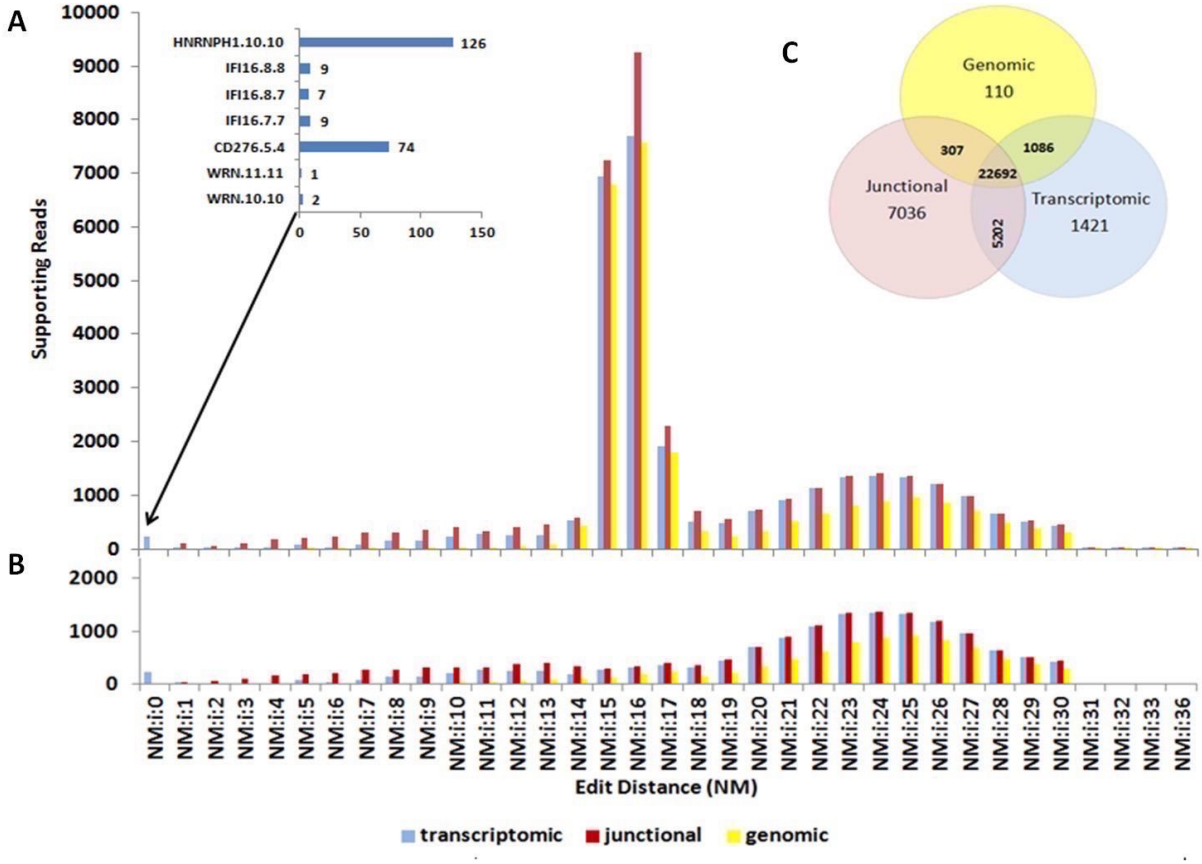


Figure 3.2. **Alignment quality of reads excluded by filters.** A) Edit distance distribution of reads filtered out by genomic, transcriptomic and junctional (JSpan/ PID) filters. Inset: Seven structures are supported by 228 reads with 100 % alignment but are excluded by the transcriptomic filter. B) 30 % of reads filtered out support a false positive structure from 5.8 s rRNA and are excluded in this plot. C) Venn diagram showing number of reads excluded by filters. Majority of false positive reads are excluded by all three filters. Each filter also excludes a distinct population of false positive reads. Figure and legend taken from Izuogu et al., (2016).

3.3.1 Filters Target Overlapping Populations of Reads

From a total of over 200 million reads (in SRR444975A), approximately 0.17% (359,837) have shuffled co-ordinates with respect to exon position, resulting in 46,875 putative PTES models. When these models were evaluated by remapping full length reads to sequence references built with information from the models, only 44,620 (~12.5%) mapped to PTES

constructs (Table 3.1), indicating that most of the reads with rearranged anchor pairs do not map to single genes and/or known exon junctions. Approximately 85% (37,854) of reads mapping to PTES constructs are removed by the genomic, transcriptomic and junctional (JSpan and PID) filters, with the majority being identified by more than one filter. For instance, over 98% of reads excluded by the genomic filter are also excluded by the transcriptomic filter, and 60% (22692) of all filtered reads are identified by all three (Fig 3.2C). Most of these have high edit distances (>10), indicative of low quality alignment. Despite this, the genomic, transcriptomic and junctional filters (at least stringency: JSpan = 4, PID = 0.60) uniquely exclude ~0.25% (110), ~3.2% (1421) and 15.8% (7036) of reads mapping to PTES models respectively, (Venn diagram, Fig 3.2C) indicating that none is wholly redundant. The subset of reads identified specifically by the junctional and transcriptomic filters are defined by low edit distances of between 1 and 10 (Fig 3.2A), and a small number of reads excluded by the transcriptome filter (228) map perfectly to putative PTES constructs with NM=0 (inset Fig 3.2A). Figure 3.2A also shows a bimodal distribution of mapping qualities for reads filtered out by all three filters with peaks at NM=16 and NM=24. Most of the excluded reads with NM=16 support a false positive structure from rRNA 5.8s (NR_003285.1.1). This is a single exon gene with more than one RefSeq annotations and without canonical splice signals. Because of its high abundance in cells, reads supporting this structure are likely to arise from template switching, hence the high number of edits in aligned reads. In Figure 3.2B, these reads are removed, altering the distribution of mapping qualities for excluded reads. This false positive structure is unlikely to be found within ribosome depleted RNAseq datasets; nevertheless, reads supporting this structure were excluded by all three filters.

	SRR444975A (Undigested)	SRR444974A (RNase R Digested)
Library Size	206362733	158305855
Reads Detected With Shuffled Coordinates - ***discovery phase***	359837	471109
Reads Mapped To PTES Models - ***evaluation phase***	44620	129347
Reads Excluded By Genomic Filter	24195	9330
Reads Excluded By Transcriptomic Filter	30401	12213
Excluded Reads with 100% Alignment to PTES	*[228]*	*[115]*
Reads Excluded By Both Genomic & Transcriptomic	*[23778]*	*[6317]*
Reads Excluded By Genomic Filter Only	*[417]*	*[3013]*
Reads Excluded By Transcriptomic Filter Only	*[6623]*	*[5896]*
Reads Excluded By Junctional Filter (PID=60, Jspan=4)	34820	32623
Reads Excluded By Junctional Filter Only (PID=60, Jspan=4)	7036	26372
Post-filter PTES supporting reads	6766	87749

Table 3.1. **Summary of excluded reads.** Number of reads excluded after each filter during PTESFinder analyses of RNase R digested and undigested samples from human fibroblasts.

3.3.2 Reads Excluded By Specific Filters Have Different Origins

To investigate the activity of specific filters further, the mapping coordinates of reads removed by the genomic filter were compared to the coordinates of annotated pseudogenes and segmental duplications. This established that ~74% of reads excluded by the genomic filter had superior alignments to segmental duplications, and ~12% had superior alignments to pseudogenes. The 417 reads identified by the genomic filter but not by the transcriptomic filter were also found to be enriched for reads derived from segmental duplications and pseudogenes (examples in Figure 3.3A).

Reads with perfect alignment to the constructs (NM=0) but excluded specifically by the transcriptomic filter, were extracted and examined further to investigate their origins. They support 7 putative PTES structures from 4 genes (Inset, Fig 3.2A). Manually examination of these 228 reads using BLAT (Kent 2002) established that they all also mapped contiguously with ~100% identity to the transcriptome due to high sequence identity between neighboring exons. For example, 126 reads which support a putative single exon PTES structure (exon 10 of *HNRNPH1* circularized) map with ~100% identity to exons 10 and 11 of the canonical *HNRNPH1* transcript (Fig 3.3B) due to high sequence identity between these neighboring exons. Another example is the putative PTES from *CD276* locus, involving exons 5 and 4. In that locus, exons 3 and 5 are ~100% identical, so are exons 4 and 6 (Fig. 3.3C-D). Reads supporting this structure map with 100% identity to the canonical junction between exons 3 & 4. As a result, these reads cannot be taken as supporting evidence for PTES. It is noteworthy that such structures will pass any qualitative filter criterion requiring only unambiguous mapping to PTES constructs, illustrating the value of the transcriptome filter. Manual examination of a subset of the 7036 reads identified only by the junctional filters established that these support structures with distinct patterns of suboptimal mapping, such as low alignment quality specific to only one of the two exons in the structure (e.g. Fig 3.4A-B), and low sequence identity specifically at the junction (e.g. Fig. 3.4C-D), the latter being consistent with the expected pattern of alignment for template switching artefacts (Houseley & Tollervey 2010).

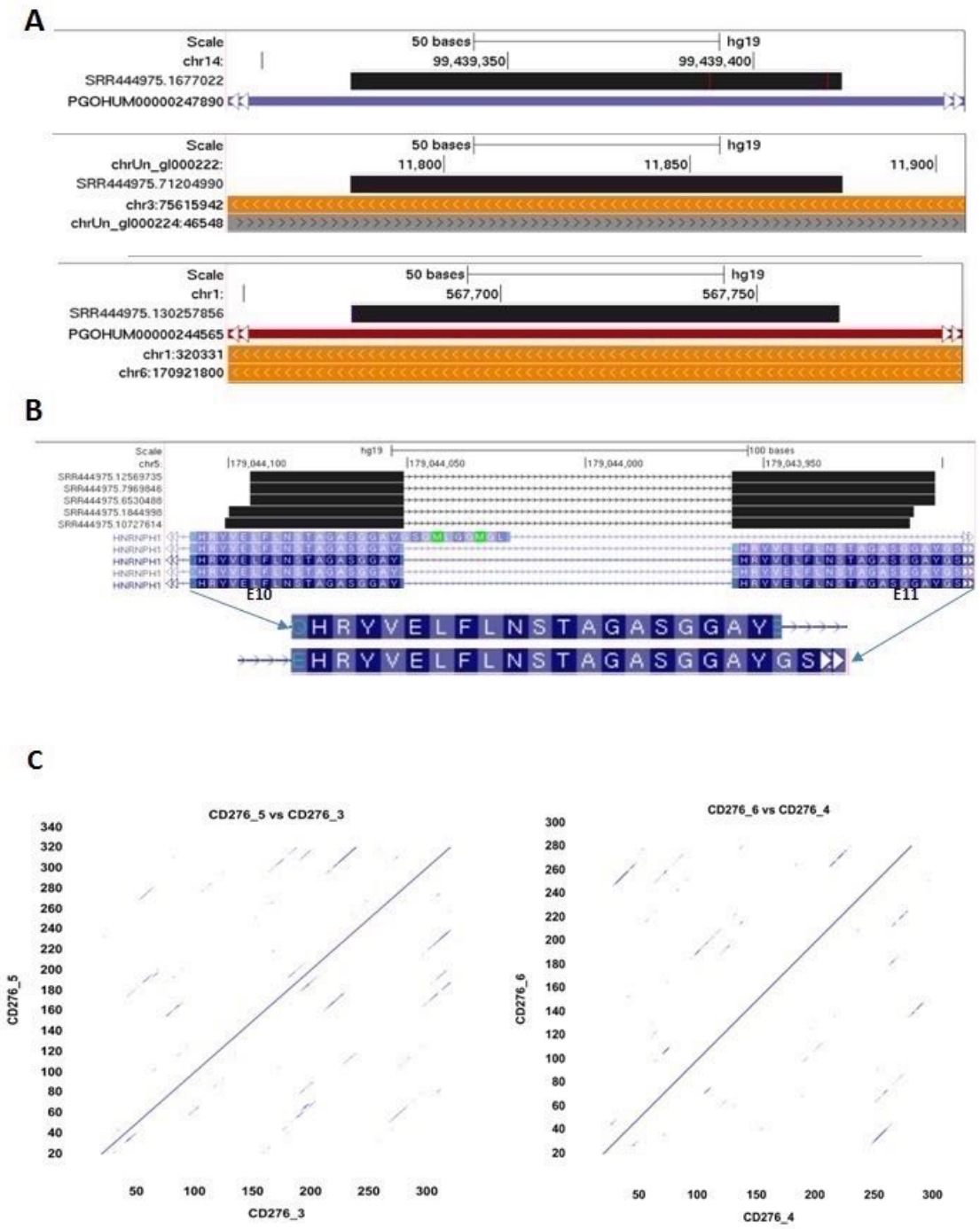


Figure 3.3. **Examples of Reads excluded by genomic and transcriptomic filters**. A)Reads filtered out by genomic filter for mapping better to pseudogenes & segment duplicated regions B) Reads excluded by the transcriptomic filter for having 100% alignment to a canonical splice between exons 10 and 11 of *HNRNPH1* C) Dot plot alignment of *CD276* exons, showing high sequence identity between adjacent exons. Figures 3.3A -B taken from Izuogu et al., (2016).

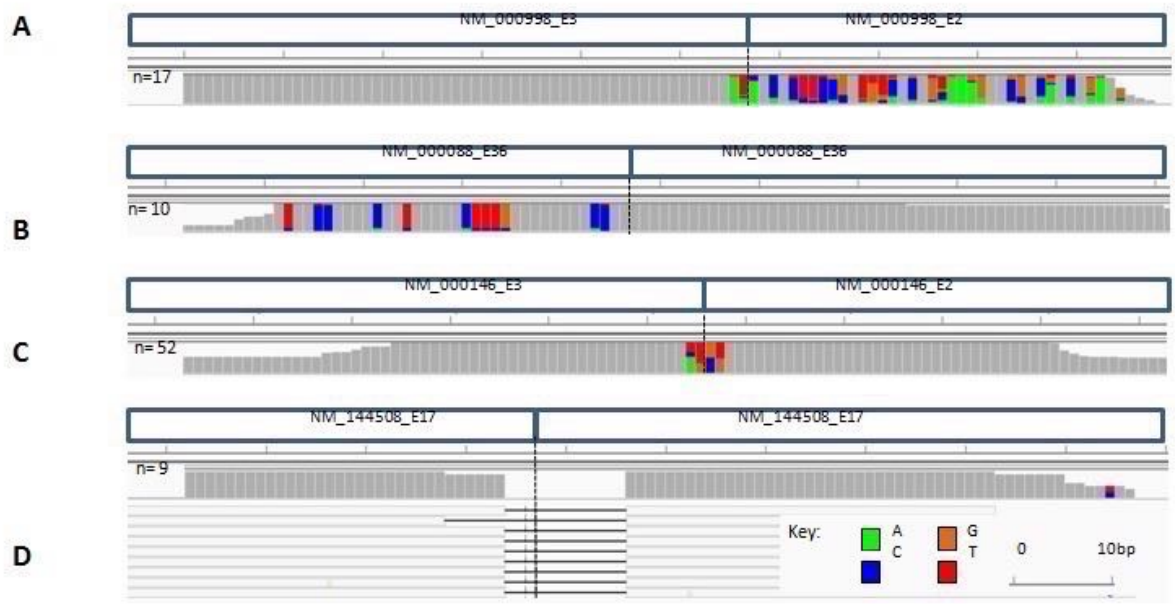


Figure 3.4. **Examples of Reads excluded by the junctional filter. A - B)** Reads excluded by minimum segment PID of 85%, mapped across putative PTES junctions. PTES junctions are denoted by line between exons; number of reads mapped across junction is shown, and mapped positions shown in grey histograms. Positions of mismatches are shown in green, blue, orange and red for A, C, G & T respectively. **C - D)** Reads excluded by JSpan, gapped and mismatch alignment around junction are apparent. Figure taken from Izuogu et al., (2016).

To further investigate the potential impact of these false positives, assessment of filters was repeated using RNAseq data derived from human fibroblast RNA which had been pre-digested with RNase R. This selectively removes linear RNAs, and enriches for circRNAs (Danan et al. 2012; Jeck et al. 2013; Jeck & Sharpless 2014), and has been shown to significantly increase the recovery of PTES reads. However, it is anticipated that this would also selectively remove false positives derived from pseudogenes and segmental duplications which mimic PTES structures, without necessarily reducing reverse transcription artefacts such as template switching. This is indeed the case, as only ~12% of reads from the RNase R digested sample which map to PTES sequence constructs are excluded by the genomic and transcriptomic filters (Table 3.1), compared to 69% in the undigested sample. Furthermore, only 17% of these map to segmental duplications, compared to 74% in the undigested sample. In contrast, the proportion of reads excluded by the junctional filters is considerably higher after RNase R digestion, consistent with expectation. Additionally, the large peak observed for reads excluded in the earlier analysis and consisting of reads from rRNA are not apparent in RNase R digested sample (Fig 3.5). This is presumably due to digestion by RNase R, thus, reducing abundance of rRNAs and the likelihood of template switching between molecules of this transcript.

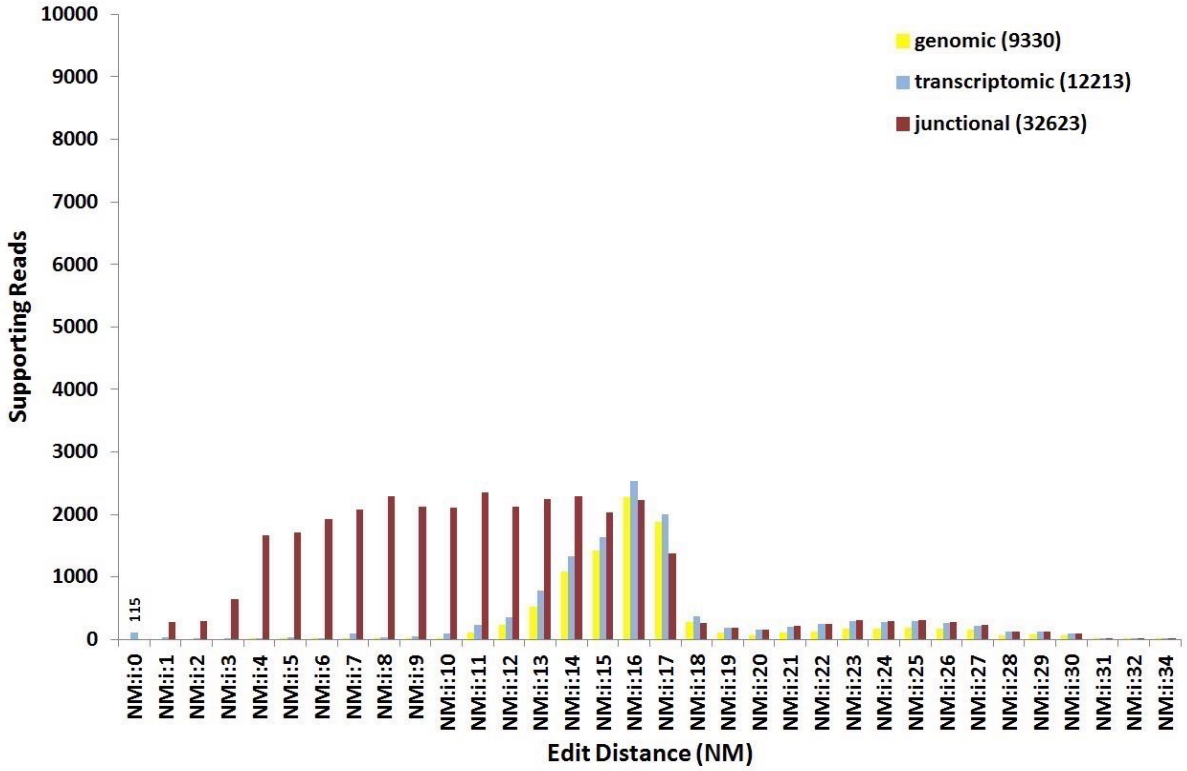


Figure 3.5. **Alignment quality of reads excluded from RNase R digested sample.** Number of reads excluded by each filter after analysis of RNase R digested sample from human fibroblasts and alignment quality distributions of reads excluded by each filter.

3.3.3 PID Has Greater Impact than JSpan

The junctional filters (JSpan and PID) examines reads aligned to constructs and require that there is no mismatch within a defined sequence region either side of the PTES junction (JSpan), and that reads map to the construct with sufficient identity above a specified threshold (PID). Reads not meeting these requirements are likely reverse transcription artefacts or originate from other chimeric transcripts not consistent with PTES. To investigate the impact of varying the user defined JSpan and PID parameters which comprise the junctional filter, the same data was re-analyzed using 54 different combinations of these parameters, both with and without the genomic and transcriptome filters applied. This established that varying the PID has a greater impact than varying the JSpan, with 5691 reads filtered with maximal PID (100%) and lowest JSpan (4) compared to only 1235 reads filtered with the maximal JSpan (14) and lowest PID (60%). Furthermore, varying the PID between 60% and 75% has little impact at any JSpan value, but above 75% there is a linear relationship with the number of reads filtered.

As the default junctional filter parameters failed to identify some reads excluded as false positives by the other filters (Genomic: 110 and Transcriptomic: 1421, Figure 3.2C above), this analysis was repeated using only these reads to establish the JSpan and PID parameters required to identify them. Over 99% of these reads are excluded with the most stringent junctional filter

parameters (Figure 3.6B). Furthermore, the vast majority are filtered with a PID of 85%, suggesting this is a logical setting for this parameter. The JSpan setting only has a major impact at low PIDs (60%-75%).

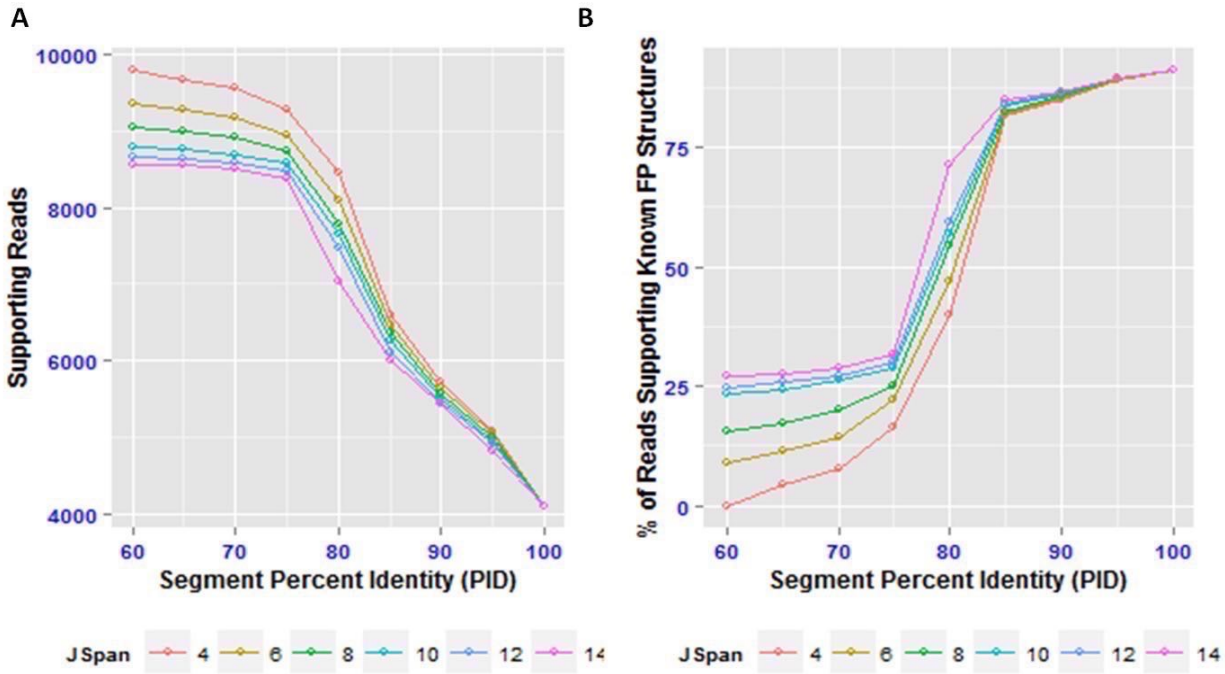


Figure 3.6. **Effect of varying junctional filter parameters**. A) Number of reads passing filter at different combinations of JSpan and PID. B) Percentage of reads only excluded by transcriptome and genomic filters at default settings, which are filtered at different combinations of JSpan and PID. Figure and legend taken from Izuogu et al., (2016).

3.3.4 Effect of Aligner Specific Parameter and PTESFinder Performance

The effect of varying the Bowtie aligner specific -m / -M toggle, which controls the uniqueness of alignments reported, was also assessed. Briefly, if the number of reportable alignments is greater than the -m value specified, all alignments for that read are suppressed. If an equivalent -M value is used, one alignment is chosen to be reported instead of complete suppression. This parameter is expected to affect the performance of PTESFinder in different ways. Accepting reads mapping to multiple targets may reduce specificity as the likelihood of inaccurate identifications based on such reads is higher. Conversely, lower values of the -m parameter can affect sensitivity. Reads mapping to exons shared by multiple linear isoforms of a PTES producing gene will be suppressed if the number of isoforms from that gene exceed the specified -m value.

To assess the effect of this parameter on PTESFinder output, I analyzed simulated data, varying -m value at 2, 4, 7 & 20, both with and without guiding with previously identified PTES transcripts. As the impact of this parameter will be influenced by read depth/coverage, simulated datasets with coverage varying from 2 to 50 were analyzed. The results are presented

in Figure 3.7 and illustrate that sensitivity varies considerably with both coverage and analysis parameters. It remains below 0.6 for all –m/-M values and analysis types at coverage of 2. Sensitivity remains below 0.70 at all depths of coverage for the –m2 analysis, due to the large number of genes within RefSeq which have multiple isoforms. However, sensitivity is >90% at coverage of 10 or higher when the default (-m7) setting is used, or when -M values were used in analyses. Specificity was observed to be over 0.99 at all depths, using any –m/-M value. Lower specificities were however observed when -M values are used instead of -m values, confirming that alignments to multiple targets may impact on specificity.

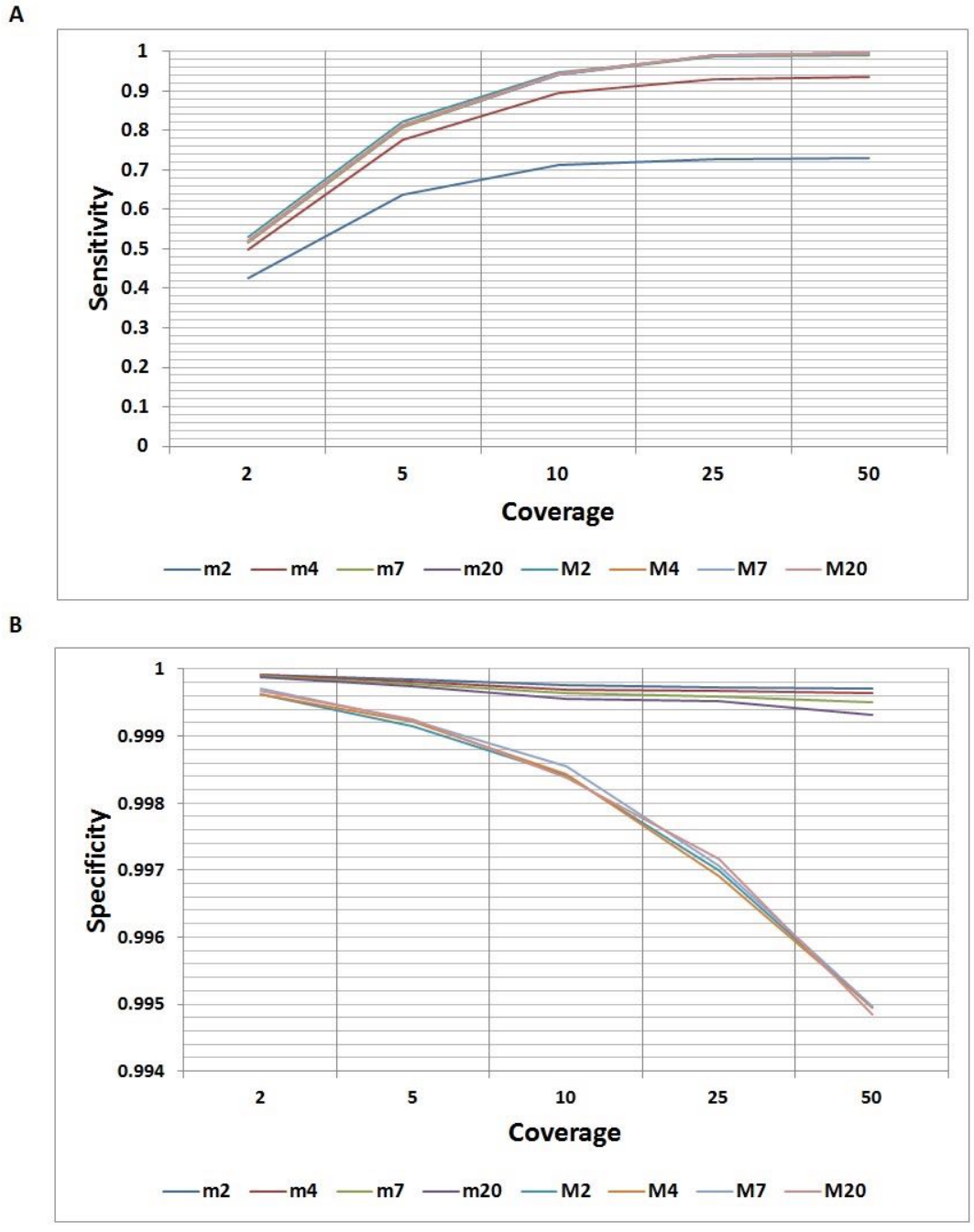


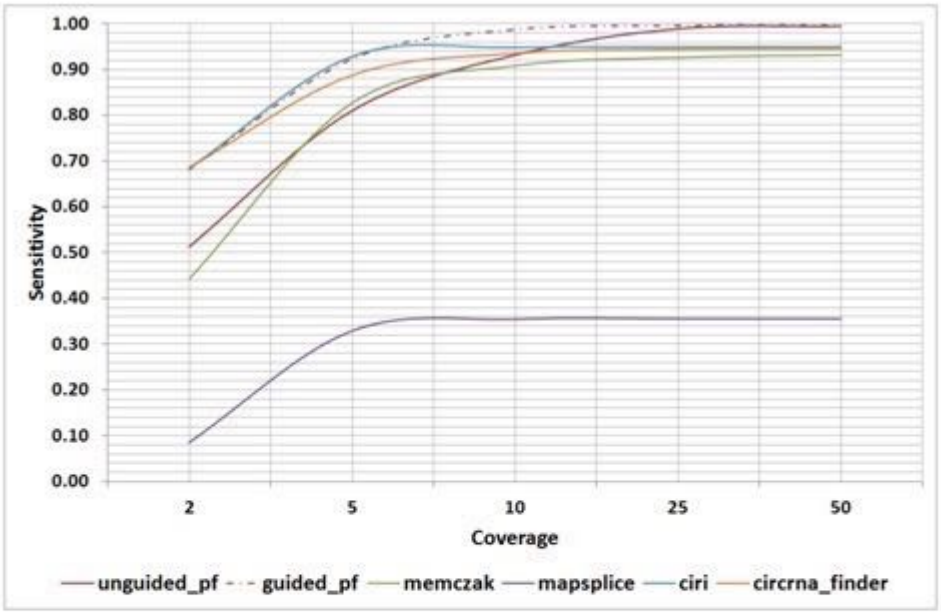
Figure 3.7. **Effect of varying aligner-specific parameter**. A) Sensitivity and B) Specificity of PTESFinder, after analyses of simulated data using various -m/M parameter values.

3.3.5 Comparison of PTES identification Methods

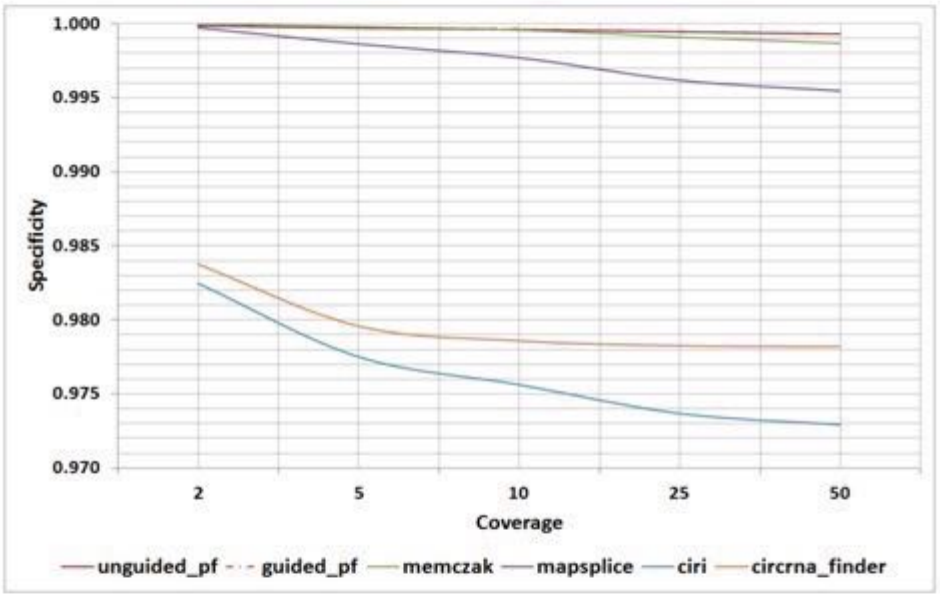
To compare PTESFinder to other methods, simulated reads were generated from previously identified PTES and associated canonical transcripts (see 2.5.2), and analyzed at various read depths of coverage using default parameters. In addition to assessing PTESFinder for *de novo* PTES discovery, the use of constructs of previously reported structures for guided discovery was also assessed, as were four publicly available methods which have previously been employed to identify circRNA transcripts: MapSplice v 2.1.5 (Wang et al., 2010) used in Jeck et al., (2013), CIRI v. 1.2 (Gao et al. 2015), circRNA_finder (Westholm et al., 2014) and the method used by Memczak *et al.* (2013). These methods utilize various established aligners in their methodologies and represent state-of-the-art tools for PTES identification. Briefly, MapSplice and the method described in Memczak et al., (2013) segment reads and map to the genome with Bowtie; CIRI utilizes BWA (Li & Durbin 2009) and circRNA_finder relies on the STAR aligner (Dobin et al., 2012). Thus, both spliced and unspliced aligners are represented as underlying mapping tools within these methods.

Results from 100 simulated datasets are presented in Figure 3.8 (A-C), and illustrate that sensitivity varies considerably with coverage, and between methods. As previously observed (Fig. 3.7), at read coverage of 2, the sensitivity of PTESFinder is below 0.6. This can be attributed to PTES junctions occurring within the terminal 20bp of reads, as the low tolerance for mismatches during anchor mapping will result in their elimination. However, sensitivity reaches >90% at coverage of 10 or higher for both guided and unguided analyses, with guided PTESFinder being equally or more sensitive than all other methods at all read depths. Strikingly, the sensitivity of MapSplice is low, remaining below 0.5 at all read depths. In contrast, specificity is over 0.97 for all methods at all read depths (Figure 3.8B), although PTESFinder achieves the highest specificities averaged across all depths (over 0.999) for both *de novo* and guided PTES discovery, with all canonical junction reads being correctly identified as such within the simulated data. Only the Memczak method has similar specificity when averaged across all read depths (Figure 3.8C).

A


1.00
0.90
0.80
0.70
0.60
0.50
0.40
0.30
0.20
0.10
0.00
Sensitivity
2
5
10
25
50
Coverage
unguided_pf
guided_pf
memczak
mapsplice
ciri
circrna_finder

B


1.000
0.995
0.990
0.985
0.980
0.975
0.970
Specificity
2
5
10
25
50
Coverage
unguided_pf
guided_pf
memczak
mapsplice
ciri
circrna_finder

C

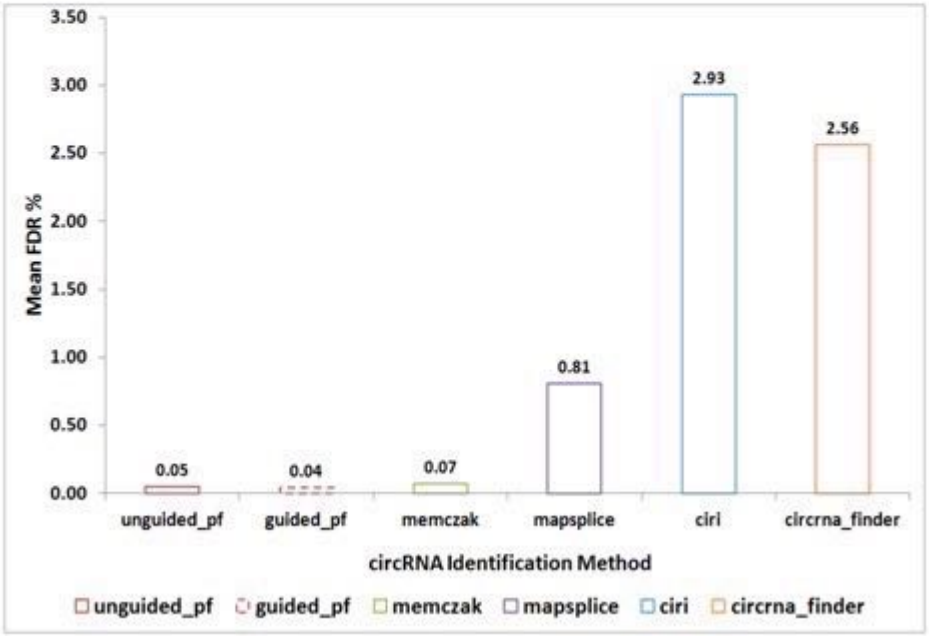

3.50
3.00
2.50
2.00
1.50
1.00
0.50
0.00
Mean FDR %
0.05
0.04
0.07
0.81
2.93
2.56
unguided_pf
guided_pf
memczak
mapsplice
ciri
circrna_finder
circRNA Identification Method
unguided_pf
guided_pf
memczak
mapsplice
ciri
circrna_finder

Figure 3.8. **Sensitivity and Specificity in Comparisons to Other Methods.** A) Sensitivity and B) Specificity of PTESFinder and 4 other publicly available methods (CIRI, circRNA_finder, MapSplice and Memczak) analyzed using simulated data (see methods). C) Mean false discovery rate (FDR) % of all methods averaged across all read depths analyzed. Figure and legend taken from Izuogu et al., (2016).

To compare performance using real data, RNAseq data from Jeck et al. (2013) were analyzed using all 5 methods (Table 3.2). To allow direct comparison to PTESFinder, the number of putative circRNA structures identified which utilise 2 RefSeq splice sites was recorded for all other methods (bracketed in Table 3.2), as the total numbers include structures from intergenic and intronic regions of the genome. For all 4 samples analyzed, CIRI consumed >90Gb of memory, resulting in incomplete analyses. It was therefore not analyzed further. Of the remaining 4 methods, PTESFinder identified on average 15% more structures than the Memczak method and ~70% more than MapSplice. The latter is consistent with earlier observation that MapSplice, which was used in Jeck et al., (2013), has low sensitivity at all depths of sequence coverage (Figure 3.8A). However, circRNA_finder reported the highest number of putative circRNA transcripts from both exonic and non-exonic regions of the genome, reporting approximately 31%-42% more structures with RefSeq co-ordinates than PTESFinder (Table 3.2).

Method		SRR444974	SRR445016	SRR444975	SRR444655
Memczak[1]	Identified	22663 (17752)	22351 (17231)	3733 (2956)	1667 (873)
	Run Time	1993m	2479m	2602m	2061m
MapSplice[1]	Identified	9701 (7087)	7380 (4891)	2231 (986)	1479 (307)
	Run Time	6167m	16356m	7412m	2605m
PTESFinder	Identified	25116	24489	5383	2316
	Run Time	1355m	1963m	1530m	1369m
circRNA_finder[1]	Identified	49901 (32856)	54154 (32186)	11069 (7309)	3130 (2131)
	Run Time	75m	90m	80m	88m

[1]circRNAs utilizing two RefSeq annotated splice sites in brackets.

Table 3.2. **Analysis of RNAseq data from human fibroblasts using 4 PTES identification tools.** Run times in minutes and number of PTES identified by each PTES identification method after analysis of RNAseq data from human fibroblasts. Table taken from Izuogu et al., (2016).

To investigate the origins of the RefSeq related structures identified exclusively by circRNA_finder, reads defining these structures from one sample (SRR444975) were re-analyzed using PTESFinder (Figure 3.9A). Of 9287 reads re-analyzed, approximately 20% (1840) are defined as mutilocus or sense-antisense fusions, and a further 19% (1775) are eliminated by the junctional, genomic, and transcriptomic filters indicating likely false positives

(Figure 3.9B). The remaining 61% (5672) are not aligned, indicating that their anchors map sub-optimally to RefSeq. Furthermore, plotting the distribution of the number of reads supporting each structure identified by circRNA_finder only, by PTESFinder only, and by both methods (Figure 3.9C), revealed that the vast majority of structures identified by circRNA_finder alone are supported by a single read. This is in sharp contrast to structures identified by both methods, or by PTESFinder alone. While these single-read structures may include *bona fide* low frequency circRNAs, they are also likely to contain false positives caused by suboptimal mapping, consistent with the lower specificity of circRNA_finder with simulated data. Runtimes and memory consumption for each method were also profiled. Runtimes for PTESFinder were 25%-35% lower than for the Memczak method, and 50%-82% lower than for MapSplice (Table 3.2), but by far the best runtimes were achieved by circRNA_finder which utilizes the STAR aligner. These were, however, achieved at higher computing memory cost (~30GB). PTESFinder and the Memczak method registered the least memory consumptions and are seemingly unaffected by increased library size.

To compare PTESFinder's output to previous reports, RNAseq data previously mined in two further studies (Salzman et al., 2012; Memczak et al., 2013) were also analyzed. Consistent with the above, it identified 13% more distinct structures from leukocyte and HEK293 data than were reported by Memczak *et al*., (2013) (2217 as opposed to 1950 Figure 3.9D), and 41.6% more structures than reported by Salzman *et al*., (2012) from leukocyte data (1875 as opposed to 1324). As both structures and supporting reads were reported by Memczak et al., (2013), it was possible to re-analyze the 898 structures identified exclusively by their method using PTESFinder. This established that none correspond to structures which PTESFinder is designed to identify (Figure 3.9E): 503 (56%) are derived from intronic, and intergenic regions, and of the 1420 reads supporting the remaining 395 genic structures, 492 were excluded by PTESFinder due to low map quality (200) or multiple map locations (292), 89 reads were excluded by PTESFinder filters, and the remaining 839 possessed at least 1 exon boundary which did not map to known splice junctions (Figure 3.9F). Again, while some of these latter reads will undoubtedly correspond to *bona fide* PTES structures (as a number of genic PTES utilizing non-Refseq splice sites have been confirmed experimentally (e.g. Memczak et al., (2013)), further BLAT (Kent 2002) analysis established that 13 mapped in a linear fashion to 6 annotated pseudogenes (see appendix 9.2).

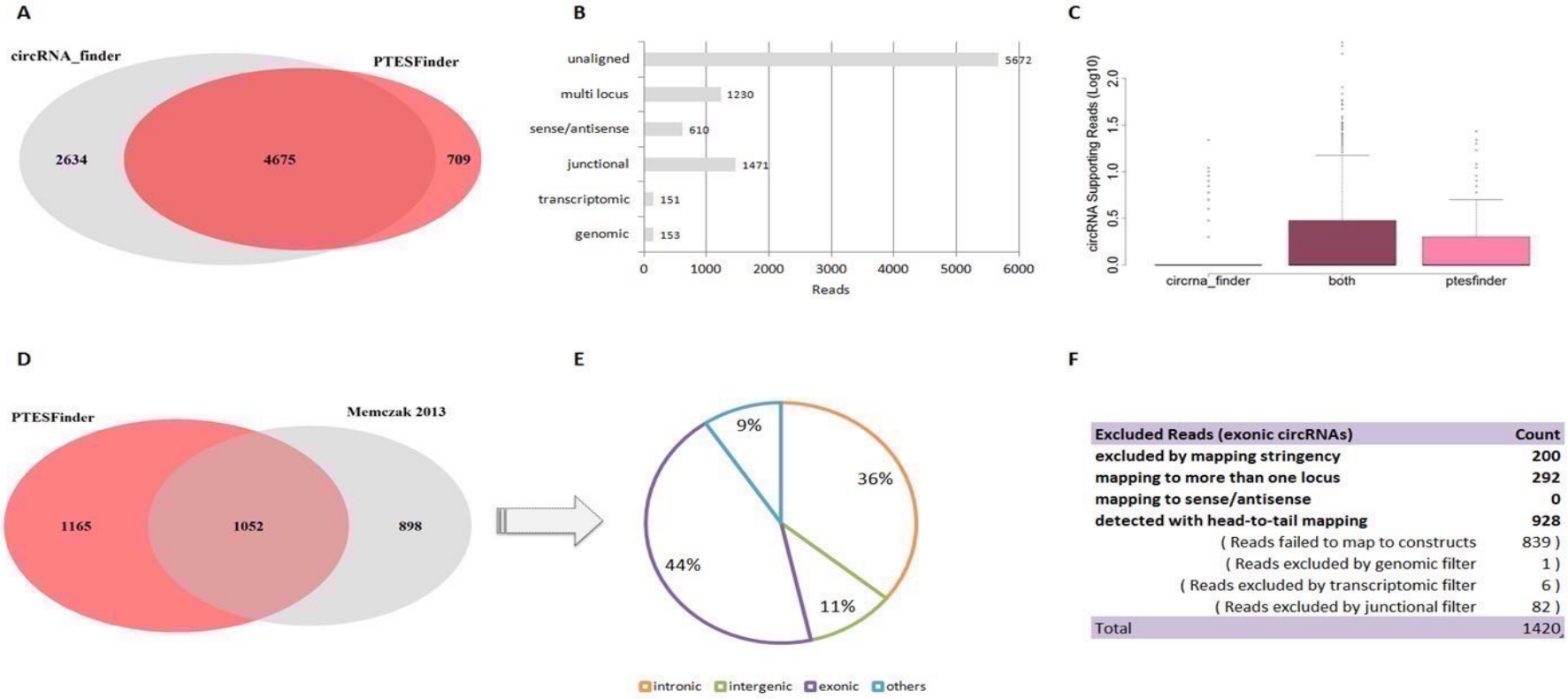


Figure 3.9. **Comparisons with real RNAseq data & published results**. A) Approximately 64 % (4675) of PTES transcripts utilising 2 RefSeq (known) splice sites were identified by both circRNA_finder and PTESFinder from SRR444975 B) Read exclusion criteria for PTES transcripts identified by circRNA_finder only, when analyzed by PTESFinder C) Distribution of read numbers supporting PTES transcripts identified by circRNA_finder only, by PTESFinder only, and by both (raw counts reported by PTESFinder shown) D) PTESFinder identified over 50 % (1052) of transcripts reported in Memczak et al., (2013). E) The majority of the 898 structures reported by Memczak et al. (2013) but not identified by PTESFinder are intronic or intergenic. F) Exclusion criteria for reads presented as evidence for exonic structures in Memczak et al., (2013) which were not reported by PTESFinder. Figure and legend taken from Izuogu et al., (2016).

3.3.6 Assessment of Annotation-Free PTES identification methods

Approaches to PTES discovery involve a compromise between the ability to detect all potentially rearranged transcripts, and the ability to identify artefacts generated as a result of the sequence and structural complexity of eukaryotic genomes, and of current library construction methods. Strikingly, the overlap between predictions is highly variable, suggesting that each method has its biases and perhaps many false positive predictions. For instance, <10% of transcripts identified from SRR444975 using the circRNA_finder method overlap with transcripts identified using MapSplice.

To better understand the effect of choice of aligners, I modified PTESFinder to allow for annotation-free identification from both exonic and non-exonic genomic regions. I developed two versions, one with STAR used in both discovery and evaluation phases (spliced method). The second version also utilizes STAR in the discovery phase but alignment of full length reads to the genome in the evaluation phase is done using Bowtie2 (unspliced method). The difference between both versions is likely to be observed in the quality of excluded reads and predicted PTES transcripts. Summarised results of output from each method after reanalyses of SRR444975 are presented in Fig. 3.10. First, more transcripts are identified from the unspliced method (11,751) than the spliced method (10,401). The unspliced method identifies virtually all transcripts also identified by the spliced method (Fig 3.10B). The main difference between the two methods is observed in the reads excluded by the filters. Expectedly, more reads are excluded by the spliced method; most of these reads were excluded by the transcriptomic filter. This is likely due to improved alignments to unannotated splice junctions, thus, lower number of edits compared to reads mapped to constructs. Typically, for structures identified from non-genic regions, supporting reads are not adequately compared to alignments to the genome, if such alignments are sub-optimal using the unspliced method. Reads support for the 1354 additional transcripts identified using the unspliced method are significantly lower (p-value < 2.2×10^{-16}, Wilcoxon rank sum test), compared to the level of support for common predictions, suggesting that many are likely to be false positives.

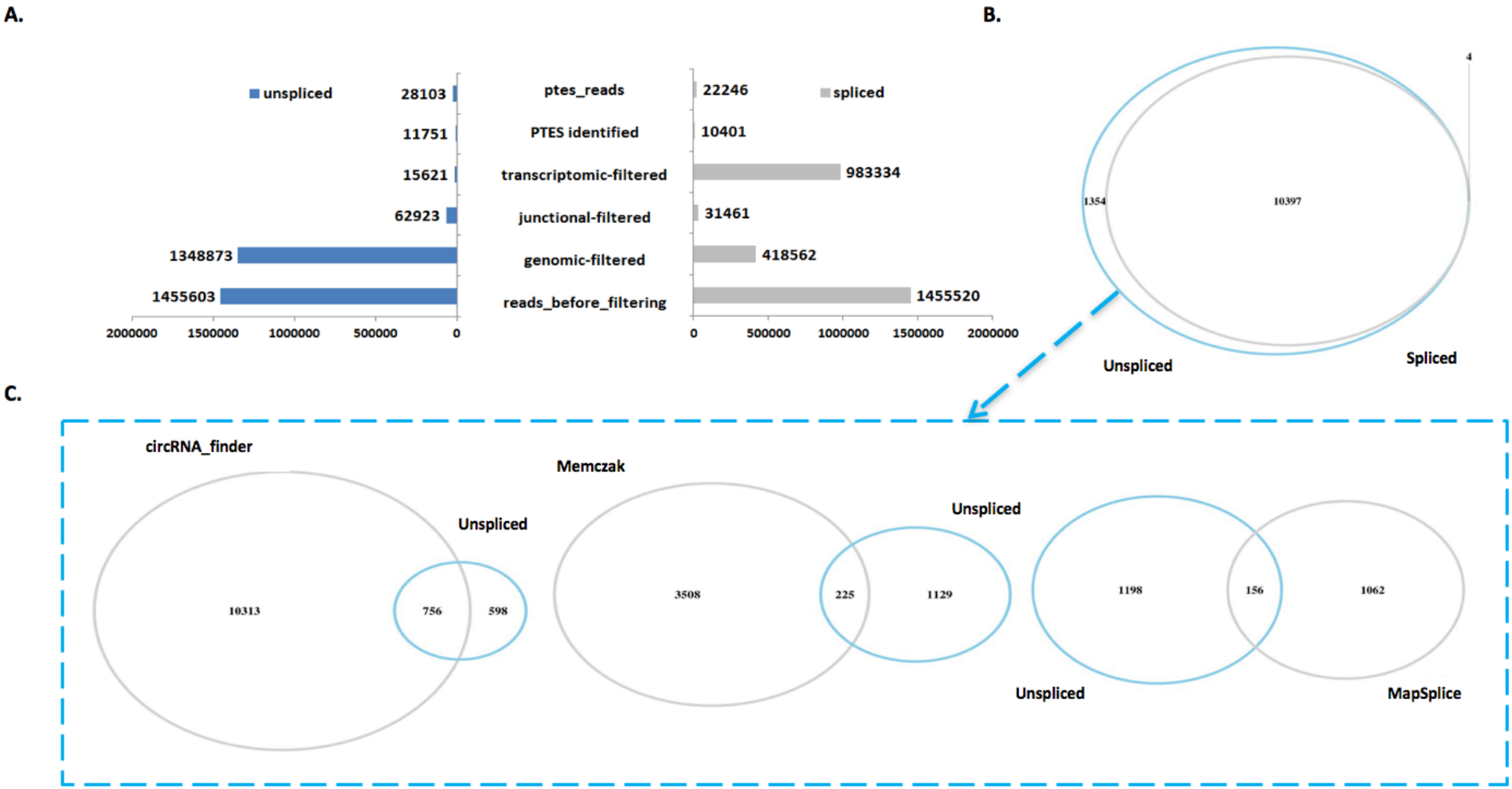


Figure 3.10. **Annotation-Free PTES identification.** A) Summary of reads excluded by both spliced and unspliced annotation-free methods. B) Overlap of transcripts identified by both methods C) PTES transcripts filtered out by the spliced method are identified by 3 published methods: CircRNA_Finder, Memczak et al., (2013) method and MapSplice.

Of the 1354 transcripts, 64 are fully intergnic and 69 transcripts span across more than one locus, such as the transcripts shown in Fig 3.1. Figure 3.11 shows an example of an intergenic transcript only identified using the unspliced method. The single read supporting this transcript has 3 mismatches when mapped to the putative PTES transcript. Interestingly, this read does not align to the genome when mapped using Bowtie2; however, an alignment is reported using the spliced aligner. STAR maps the read to an unassigned contig of the genome in a split read manner, presumably as a result of an unknown splice junction or RT-PCR artefact. As this read is not mapped to the genome in the unspliced method, the suboptimal alignment to the PTES transcript is accepted; this is not the case using the spliced method. The reported alignment to the genome in the spliced method is superior to alignment to the PTES transcripts, resulting in the exclusion of that read. It follows that methods that identify PTES from reads not mapped to the genome using non-splice aware aligners are likely to wrongly characterize structures such as this as *bona fide* PTES transcripts.

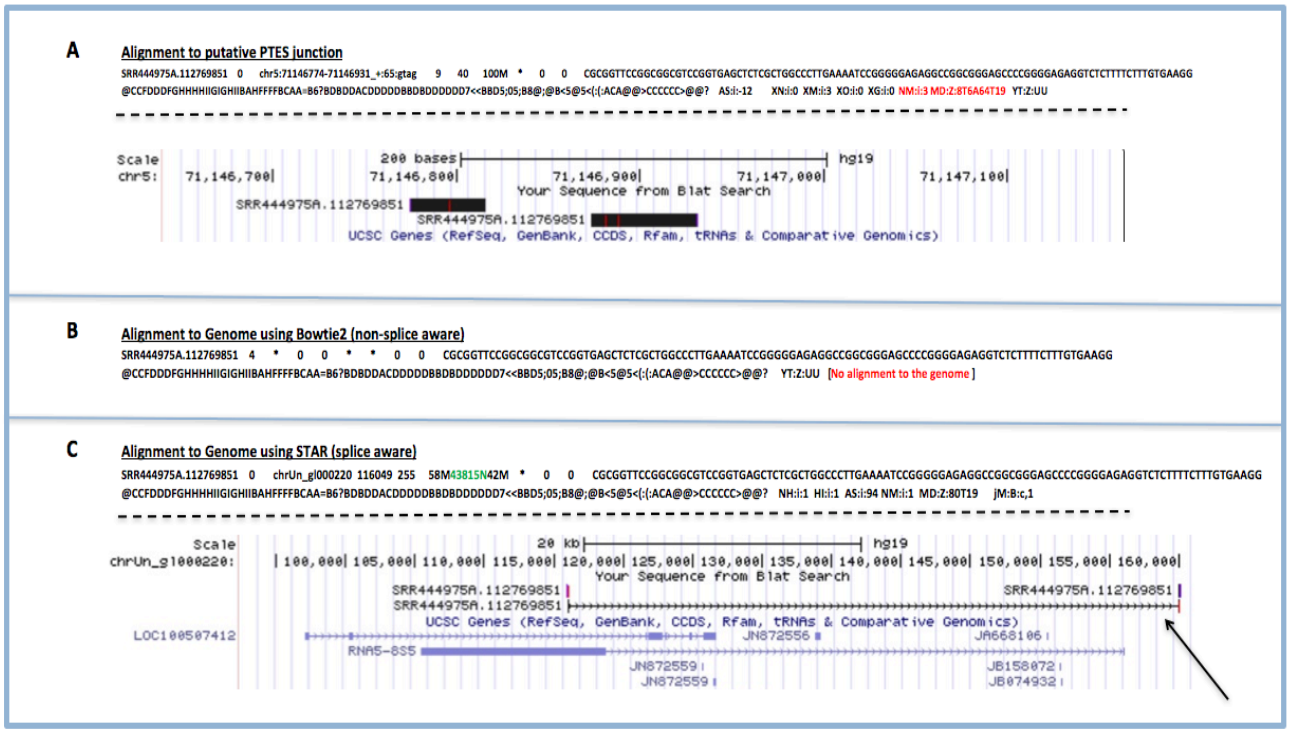


Figure 3.11. **Example of structure excluded by spliced-aware PTESFinder. A**) Example read alignment summary to a putative PTES junction. Alignment summary (in SAM format [https://samtools.github.io/hts-specs/SAMv1.pdf]) includes read id (first column), target mapped to (third column) and alignment quality summary indicated by NM (number of edits necessary for reported alignment) and MD fields – both highlighted in red and indicate indicate suboptimal alignment to the putative PTES. Lower panel shows the alignment on the UCSC genome browser (Kent 2002); two black boxes represent non-contiguous (head-to-tail) alignment of both ends of the read. **B**) Alignment summary of same read when mapped with Bowtie2, non-splice aware aligner used in the unspliced method. Missing target in third column of the alignment summary indicates that the read was unaligned to the genome using this method. **C)** An alignment is howver reported for this read using STAR, a splice-aware

aligner used in the spliced method, resulting in the correct exclusion of this read during filtering. Cigar field (highlighted in green) shows the size of the gap between both ends of aligned fragments. The arrow highlights the same alignment viewed on the UCSC genome browser and indicates mapping to an unannotated region of the genome (chrUn_g1000220).

A total of 756, 225 and 156 structures predicted by CircRNA_finder, Memczak method and MapSplice respectively overlap with the 1354 excluded by the spliced method. The lower overlap with MapSplice and Memczak methods is likely due to overall lower number of identified transcripts using these methods. From the -m/M value analysis, it is clear that there is also an aligner-specific behavior in identification of reads with head-to-tail mapping. This is likely to be the case for the higher concordance between circRNA_finder predictions and transcripts only identified using the unspliced method, as both methods use identical aligner-specific parameter values. It also further highlights the possibility that a high proportion of transcripts identified by circRNA_finder are likely false positive structures.

3.4 Discussion

To fully characterize PTES transcripts transcriptome-wide, an accurate computational method for identification is paramount. The major challenge in *in silico* identification of these transcripts is discriminating between *bona fide* PTES transcripts and all known sources of artefacts. In this chapter, I assessed the efficacy of filters within PTESFinder designed to minimize false positive predictions. Many existing tools for PTES discovery initially map reads to the genome to eliminate reads with known origins. Reads with unknown origins are often processed further to identify PTES events. PTESFinder approaches this task differently: All reads are used in PTES identification prior to filtering, where alignments to putative PTES transcripts are compared to alignments of same reads to the genome and transcriptome. This approach enhances discovery, as reads that support PTES may inadvertently be excluded by forced alignments to the genome using other methods.

Filters within PTESFinder target an overlapping population of false positive reads but have unique populations of excluded reads. Reads excluded by the genomic filter have better alignments to segment duplications and unspliced pseudogenes, than to putative PTES transcripts. Similarly, some reads excluded only by the transcriptomic filter originate from canonical splices between exons with high sequence identity to adjacent exons. These reads are most likely to be wrongly mischaracterized at PTES supporting reads because they map 100% to putative PTES transcripts. The importance of this filter provides justification for reliance on curated transcript annotations to guide discovery. It is now clear that the majority of transcripts with re-arranged exon order utilize known exon junctions (Jeck et al., 2013; Liang & Wilusz, 2014) which are processed by the spliceosome (Guo et al., 2014; Ashwal-Fluss et al., 2014). As a result, methods which utilize existing transcript annotations from the genome under study, such as PTESFinder and those employed by (Salzman et al., 2012 & 2013) benefit from the reduced noise inherent in this approach and are suited to quantitative analyses of PTES structures that can be characterized using existing annotations. The use of known/experimentally verified splice sites does reduce the misidentification of template switching artefacts or unspliced pseudogenes as *bona fide* PTES transcripts. However, it does mean that not all rearranged transcripts will be identified. All the methods compared with PTESFinder attempt to detect PTES events in non-exonic regions.

The trade-off between sensitivity and specificity in PTES identification is highlighted when the choice of aligner in annotation-free methods is considered. The discordance between PTES identified with different methods from the same samples have been reported (Yu et al., 2014; Hansen et al., 2015). This discordance is likely due to the underlying aligner used in each method and aligner specific parameters. Assessing the effect of aligner choice on PTES

identification, approximately 11% of transcripts identified when a non-splice aware aligner is used are excluded when the same sample is analyzed with a method utilising a splice-aware aligner.

The transcripts not discovered using a splice-aware aligner are likely false positives, excluded because of their similarity to canonically-spliced transcripts. It is however apparent that a large number of transcripts identified using annotation-free methods are missed by the annotation-dependent method. Many of these transcripts originate from cryptic splice sites within (or flanking) annotated genes, as is the case for circSry (Capel et al., 1993); or from antisense transcripts (eg. circCDR1 [Hansen et al., 2011; Memczak et al., 2013]); or intergenic and unannotated. It follows that the choice of PTES identification tool will depend on underlying research objectives. While annotation-dependent tools (like PTESFinder) are best suited for identifying exonic circRNAs, linear PTES from known genes and estimating their abundance relative to cognate canonical junctions; annotation-free methods will be best suited for identification of PTES events from unannotated regions of the genome. Nevertheless, comparing PTESFinder to other published methods using both simulated data and publicly available RNAseq datasets, PTESFinder achieves the highest specificity and comparable sensitivities at all read depths tested. Experimental validation of 40 out of 45 randomly selected PTES transcripts performed during my MRes were consistent with results from performance tests obtained here; further suggesting that PTESFinder is an adequate computational tool for PTES discovery.

Finally, it is notable that *in vitro* enrichment protocols may impact computational PTES identification. CircRNA enrichment by RNase R digestion apparently yields more identified transcripts, regardless of software used. However, some false positive predictions are likely from samples not depleted of ribosomal RNA as exemplified by putative PTES from rRNA 5.8s (see text). Therefore, it is necessary to account for these potential confounding factors in experimental design for PTES analysis.

3.5 Conclusion

In this chapter, I assessed PTESFinder, a computational tool for PTES identifying, developed during my MRes study. By screening for PTES events in RNAseq data from human Fibroblasts and characterizing reads excluded by filters within PTESFinder, I established the effectiveness of these filters and their usefulness. Using simulated data, I assessed the performance of PTESFinder and the effect of aligner-specific parameters. I further compared the output of PTESFinder to that of 4 published methods after analyzing simulated data. These analyses established that PTESFinder achieves the highest specificity and comparable

sensitivity to other methods, indicating that it is an adequate computational tool for PTES identification. I also extended PTESFinder to identify non-exonic PTES transcripts, and found that some predictions are aligner specific and probably contribute to the variation in reported identifications using various methods. With the utility of PTESFinder established, I turned my attention to questions pertaining to the distribution, formation and potential function of these novel transcripts.

Chapter 4. Sub-Cellular Distribution of PTES transcripts and their contribution to the Proteome

4.1 Introduction

In the last chapter I presented an assessment of PTESFinder, a computational tool for PTES discovery. PTESFinder was shown to achieve the highest specificity and comparable sensitivity when compared with other published methods. Having established the utility of this tool, I aimed to address questions pertaining to the origin and potential functions of these novel transcripts. One such question relates to defining the sub-cellular distribution of PTES transcripts. RNAs typically localize in their site of functional relevance. For protein coding genes, transcripts are transcribed and processed in the nucleus, then exported to the cytosol where they are translated by polysomes. Small RNAs, including snoRNAs and snRNAs, are generally localized in the nucleus (Elliott & Ladomery, 2011). For instance, snoRNAs are transported to the nucleolus (after transcription in the nucleoplasm), where they aid in the maturation of ribosomal RNAs. They are typically enriched in the nucleus and not in the cytosol (van Heesch et al., 2014). Similarly, studies have concluded that a subset of lncRNAs (including *NEAT1 & MALAT1*) is enriched in the nucleus, where they are functionally relevant (Derrien et al. 2012; van Heesch et al. 2014).

4.1.1 Spliceosomal proteins aid nucleo-cytoplasmic mRNA export

There are various pathways for transcripts to exit the nucleus into the cytoplasm. For spliced protein-coding transcripts, serine-arginine (SR) proteins associate with nascent transcripts prior to splicing and aid in spliceosome assembly (Elliott & Ladomery, 2011). Along with SR proteins, other proteins are added to the transcript including the cap-binding complex (CBC), exon-junction complex (EJC) and proteins associated with polyA-tail formation. These proteins coat the transcript and interact with export adaptor proteins (TREX complex) (Masuda et al. 2005). First, RNA export factor (within TREX) associates with CBC. Second, SR proteins, which help in spliceosome assembly when phosphorylated, dephosphorylate to allow binding to export adaptors (Köhler & Hurt 2007; Elliott & Ladomery, 2011). Interestingly, dephosphorylation of SR proteins acts as an indicator for spliced and unspliced transcripts; whereby unspliced transcripts are not export competent and remain in the nucleus (Elliott & Ladomery, 2011). Export adaptors are required to bind to export receptors - called Tip

associated protein (TAP) - on nuclear pores. TAP has both hydrophobic and hydrophilic termini that allow it to bind TREX and aid the movement of spliced transcripts through nuclear pores (Elliot & Ladomery, 2011). Unlike mRNAs, non-coding transcripts (such as tRNAs, rRNAs and miRNAs) utilize different export pathways (Fig. 4.1) (Kohler & Hurt, 2007; Elliott & Ladomery, 2011). Generally, these ncRNAs bind to cargo proteins that shuttle between cellular compartments, facilitated by a group of nuclear export receptors called karyopherins (Elliott & Ladomery, 2011).

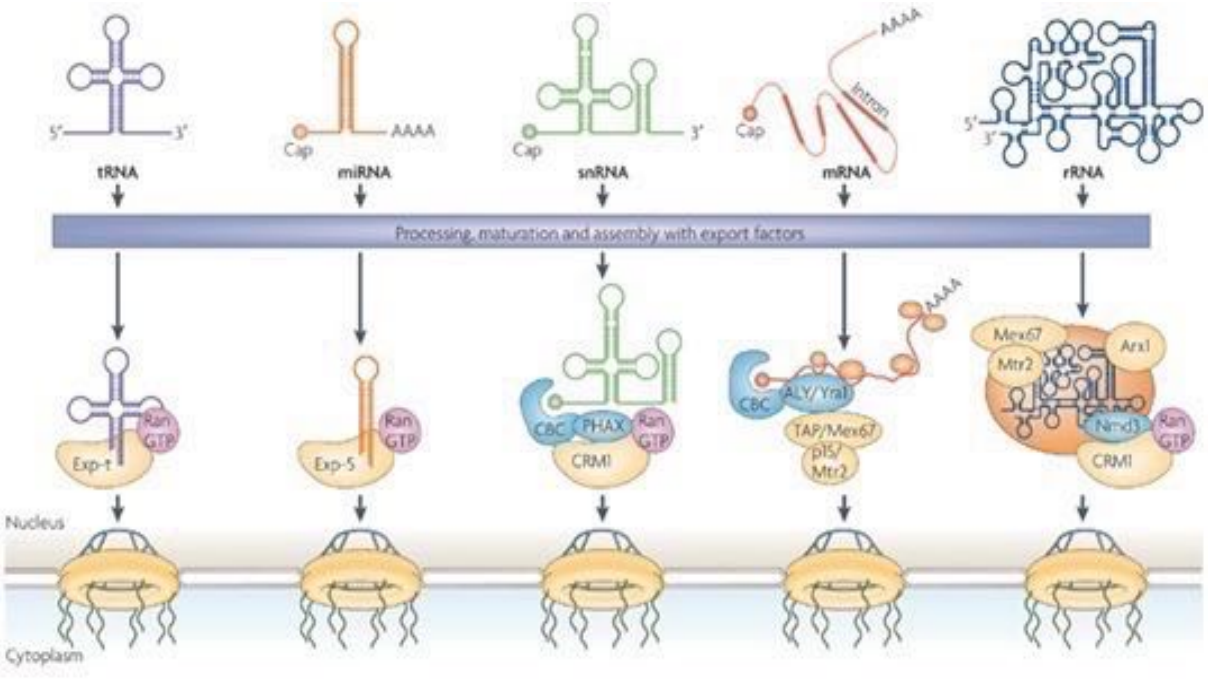


Figure 4.1. **Various RNA export pathways.** Protein-coding and non-coding RNAs exit the nucleus to the cytoplasm via various export pathways and facilitated by numerous export proteins - adapters and receptors. Taken from Kohler & Hurt, (2007).

Linear PTES transcripts can conceivably be exported from the nucleus in the same way as processed mRNA transcripts, if they are not subjected to nonsense mediated decay (NMD). This is indeed the case for some trans-spliced transcripts with detectable protein products (Caudevilla et al. 1998). However, with the absence of cap structures and polyA-tails, it is not clear if circRNAs exit the nucleus and by what mechanism. Although earlier studies had shown a handful of PTES transcripts to be localized in the cytosol, using low throughput *in vitro* methods (Nigro et al., 1991; Salzman et al., 2012), it remains unclear whether this is a transcriptome-wide property of PTES transcripts. It has been speculated that PTES transcripts may exit the nucleus during mitosis (Jeck et al., 2013). Identifying PTES transcripts localized (or enriched) in the nucleus, relative to the cytosol, would provide evidence for an unknown nucleo-cytoplasmic export pathway.

4.1.2 Co- or Post-Transcriptional Exon Shuffling?

It is becoming increasingly clear that most splicing and RNA processing occurs during transcription. Recent estimates of the proportion of exons spliced during transcription exceeds 71% (Tilgner et al. 2012), with some estimating higher co-transcriptional splicing rates of up to 80% (Girard et al. 2012). Non-coding RNAs are thought to be inefficiently spliced and have lower rates of co-transcriptional splicing (Tilgner et al., 2012). As PTES transcripts originate from both protein coding and non-coding genes, it remains unclear whether PTES biogenesis occurs during transcription and to what extent it impacts linear transcripts from the same locus. Pieces of evidence supporting both co- and post-transcriptional PTES biogenesis have recently been presented. The evidence for co-transcriptional splicing involves the identification of PTES transcripts within chromatin-associated RNA in Drosophila fly heads and mouse liver (Ashwal-Fluss et al. 2014). This approach alone does not conclusively indicate co-transcriptional PTES, as contamination of RNA extracts can occur during isolation protocols. In support of this premise, reduced numbers of circRNAs were reported from chromatin-associated samples and these transcripts are supported by lower read counts, possibly indicating contamination (Ashwal-Fluss et al. 2014). Additionally, studies have shown that many non-coding RNAs have roles in chromatin structure and that RNAs are integral components of chromatin (Mondal et al. 2010). A sub-class of circRNAs was also recently reported to associate with RNA polymerase and possibly have roles in regulating transcription (Li et al. 2015).

Evidence for post-transcriptional exon re-arrangement was recently presented by Liang and Wilusz (2014). In that study, circularization of mini-gene constructs was found to require the formation of 3` ends of the pre-mRNA, presumably to stabilize the transcript or aid formation of RNA secondary structures favorable for PTES. Removing signals for polyadenylation abolished circularization; replacing these signals by inducing the formation of a tRNA-like secondary structure that is cleaved by RNase P rescued circularization, supporting the notion that termination of transcription may be critical for PTES biogenesis (Liang and Wilusz 2014). However, it is not clear if these observations extend to other transcripts. Indeed, a follow-up study by the same group found that, biogenesis of circRNA from the *laccase2* gene in Drosophila does not require the formation of a 3` end (Kramer et al., 2015). No direct assessment of the co-transcriptional splicing profiles of PTES producing genes has been reported. Identifying a propensity for PTES genes to undergo more or less co-transcriptional splicing relative to other genes could offer further clues about how their biogenesis relates to that of linear transcripts.

4.*1.3 Profiling transcripts undergoing translation*

Once exported to the cytosol, protein-coding transcripts are bound by ribosomes and undergo translation. To identify transcripts undergoing translation, ribosome-associated transcripts are isolated and sequenced using two main approaches: 1) transcriptome-wide polysome profiling and 2) ribosome foot-print profiling (Ingolia 2014). In polysome profiling, transcripts bound by polysomes are isolated by ultracentrifugation, resulting in different fractions that indicate the number of ribosomes bound to transcripts (Ingolia 2014). In ribosome foot-print profiling, cells are first treated with cycloheximide to stall and stabilize ribosomes on mRNAs. Messenger RNAs are then digested by nuclease, but protected at regions occupied by bound ribosomes. Protected fragments are then sequenced to obtain a snapshot of specific regions undergoing translation (Ingolia 2014; Mcmanus et al. 2014).

In most organisms, protein-coding genes are characterized by open reading frames (ORFs), with translation starting at the first AUG codon and terminating at one of three stop codons (Ingolia et al. 2011). However, recent studies now indicate that there may be exceptions. There is evidence that there are internal ribosome entry sites, non-AUG translation initiation points and existence of short ORFs producing micro-peptides with little known functional relevance (Ingolia et al. 2011; Andrews & Rothnagel 2014). Indeed, ribosome profiling transcriptome studies revealed that, in some cases long intergenic ncRNAs are bound by polysomes, although these interactions may not result in functional proteins (Banfai 2012; Guttman et al. 2013; van Heesch et al. 2014).

There are conflicting reports of the potential of PTES transcripts to contribute to the proteome. Caudevilla et al., (1998) reported the identification of a linear PTES transcript from Carnitine octanoyltransferase (*COT*) locus in rat liver that produced detectable protein products. Additionally, some PTES transcripts identified in previous studies contain ORFs (Al-Balool et al. 2011; Dixon et al. 2007). As these transcripts have been observed in the cytosol, the presence of ORFs potentially raises the possibility that they may contribute to the proteome. It has also been shown that artificial circular products can be translated (Perriman & Ares 1998; Wang & Wang 2014). By attaching inverted repeat sequences and an internal ribosome entry site to green fluorescent protein (*GFP)*, Wang & Wang (2014) circularized *GFP in vivo* and subsequently observed the protein product. However, Jeck et al., (2013) investigated the association of 3 abundant circRNAs with monosomes and found no evidence that those circRNAs are bound by monosomes or are translated.

In this chapter, I assessed the distribution of PTES transcripts in the nucleus, cytosol, nucleoplasm and nucleolus of various cell lines. Identifying transcripts enriched in one nuclear compartment (and not in others) may give clues to their functional relevance or point to different mechanistic origin. To investigate whether PTES occurs co-transcriptionally, I assessed the distribution of PTES transcripts in chromatin-associated RNA and estimated the rate of co-transcriptional splicing of transcripts from their host genes (relative to other genes), using *in silico* methods. Furthermore, I investigated the protein coding potential of PTES transcripts by assessing their distribution in RNAseq data from sucrose gradient fractions of HEK293 cells, with or without arsenite treatment to inhibit translation.

4.2 Aims

In this chapter, my specific aims were:

- To assess the distribution of PTES transcripts in cellular compartments and identify PTES specific to each cellular compartment.
- To assess the extent of co-transcriptional PTES biogenesis.
- To investigate the protein coding potential of PTES transcripts.

4.3 Results

As part of the ENCODE project, RNA extracts (comprised of transcripts with size > 200bp) from cellular compartments of various cell lines were sequenced to establish the distribution of RNA species within each compartment. This, ostensibly, is the first step to inferring functional relevance of RNAs with unknown functions. For my study, I repurposed 29 RNAseq samples from that project to assess the distribution of PTES transcripts in various cellular compartments and explore any possible impact of PTES biogenesis on expression levels of cognate canonical transcripts within the cytosol. In total, ~5.1 billion reads were screened for PTES events using PTESFinder v. 1 and guided by previously reported PTES transcripts (n = 40594). This resulted in the identification of 27712 distinct PTES transcripts from 7067 genes (protein-coding and non-coding). Cumulatively, 200,593 PTES supporting reads were observed from cytosolic and nuclear samples. This represents an average of 7 reads per distinct PTES junction. Similarly, 47,784,631 reads mapped to 223,819 distinct canonical junctions, resulting in a mean of ~213 reads per distinct canonical junction. PTES transcripts contribute to the number of reads mapped to canonical junctions to varying degrees. CircRNAs, for instance, can have n - 1 canonical junctions; where n is the number of constituting exons. The lowest number of PTES transcripts were identified from NHEK cells, with 1117 and 2192 distinct transcripts from both nucleus and cytosol samples respectively. The highest number of distinct PTES transcripts (6140 and 3660) was observed from K562 nucleus and cytosolic samples respectively. This large difference in number of PTES transcripts may be due to multiple factors, including sampling bias and tissue specificity. There is also a noticeable variation in nucleus/cytosol ratio of identified transcripts, ranging from 0.31 for H1 embryonic stem cells (ESC) to 1.9 for NHEK. Full lists of identified PTES from each sample analyzed are in appendix 9.3.

In initial exploratory analysis, correlations of PTES and canonical junction counts between cell lines were assessed. Hierarchical clustering of samples was performed, based on Euclidean distance between samples. First, nuclear samples are noticeably highly correlated between cell lines, with correlation coefficient > 0.7 for some pairwise comparisons using PTES junction counts (Fig 4.2A). Second, samples seemingly group according to cellular compartments using hierarchical clustering of PTES expression. Clustering by canonical junction counts seemingly partitions samples according to cell type, regardless of cellular compartment. These results suggest that there is a detectable difference in PTES distribution between compartments, However, the observed pattern in Fig 4.2A is strongly affected by highly expressed PTES transcripts from snoRNA loci (see appendix 9.3). Both PTES and canonical junctions counts are more prominent in nuclear fractions than the cytosol (Fig. 4.2B).

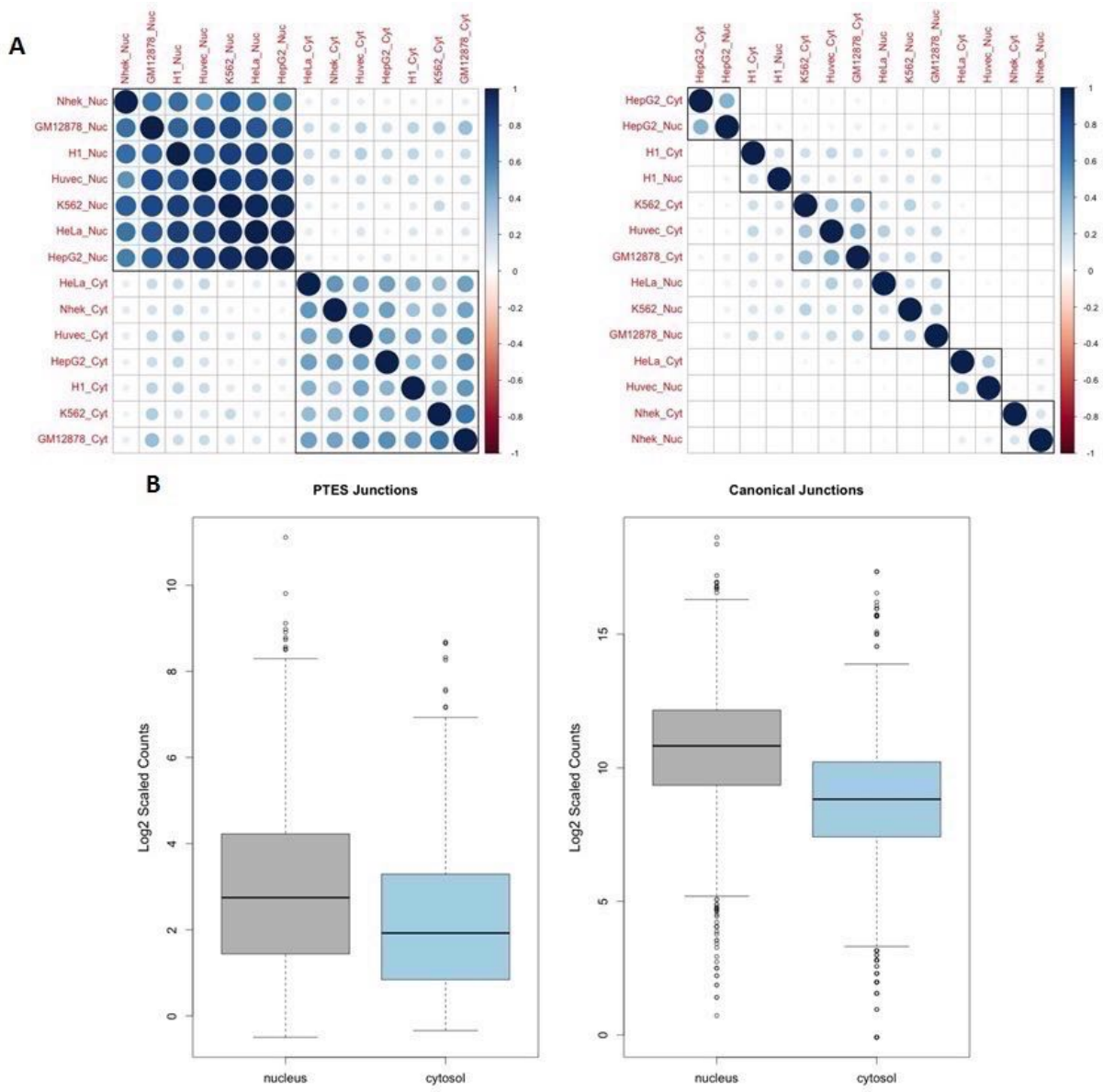


Figure 4.2. **Exploratory analysis of PTES sub-cellular distribution**. A) Correlation plots of samples using raw PTES counts (left) and canonical junction counts (right). Hierarchical clustering is based on Euclidean distance between samples. B) Distribution of log scaled and mean centered junction raw counts.

4.3.1 Variety of PTES events observed in the nucleus

To reduce the effect of potential sampling bias, transcripts identified from nuclear and cytosolic samples were post filtered to exclude transcripts observed in only one sample. Excluded transcripts were examined to ascertain whether they are over-represented in cell lines with abnormal karyotypes (cancers), which may suggest a link between PTES events and chromosomal abnormalities. About 59% of these transcripts were from 11 (of 23) samples with abnormal karyotypes, irrespective of their cellular compartment. This observation is however not statistically significant ($p = 0.3808$, Chi-square test, 1 degree of freedom). Post-filtering reduced the number of transcripts to 11491, 8477 of which were observed in both cellular

compartments (Fig. 2.2). However, 2178 transcripts were observed in the nucleus only, ~2.6X more than observed exclusively in the cytosol (836). This observation is consistent with all transcription (and PTES biogenesis) occurring in the nucleus prior to export to the cytoplasm. The median number of reads supporting transcripts exclusively identified in the cytosol is 2, lower than observed for nucleus-only transcripts (median: 3; p-value < 2.2×10^{-16}, Wilcoxon rank sum test) . This observation may suggest that identification of cytosol-only transcripts are lowly expressed and not easily be detectable in the nucleus, due to sampling.

4.3.2 snoRNA-PTES transcripts are likely artefacts

PTES transcripts only identified from the nucleus include 51 transcripts arising from single exon genes, 41 of these emanate from snoRNA genes (Table 4.1). Majority of PTES transcripts are from multi-exon genes which are spliced. However, two well characterized circRNAs from *Sry* and *CDR1* loci have been reported (Capel et al., 1993; Hansen et al., 2011; Memczak et al., 2013), but biogenesis of these circRNAs depend on cryptic splice sites outside the coding sequence of the genes (Dubin et al. 1995). They are also flanked by inverted complementary repeats that facilitate circularisation.

Human snoRNAs have sizes ranging from ~50bp to 200bp (Taft et al., 2009), are processed by nucleases and do not undergo splicing. Most are not expected to be within the samples analyzed, since only long RNAs with sizes >200bp were captured and sequenced. Strikingly, these 41 snoRNA-PTES transcripts are highly expressed and were observed in various nuclear RNA samples. Manual examination of reads supporting these transcripts showed that most of the reads aligned to the PTES models with no mismatches.

chromosome	start	stop	structure	total_counts	strand	host_gene
chr1	173835105	173835166	NR_002750.1.1	27	-	SNORD44
chr1	175937532	175937676	NR_002998.1.1	46	-	SCARNA3
chr1	45242163	45242261	NR_000024.1.1	4	+	SNORD46
chr20	37058309	37058447	NR_003017.1.1	2	-	SNORA71C
chr7	45024976	45025109	NR_002952.1.1	16	-	SNORA9
chr13	45911614	45911744	NR_002967.1.1	3	-	SNORA31
chr1	235291117	235291252	NR_002956.1.1	4	-	SNORA14B
chr11	62622763	62622838	NR_002564.1.1	245	-	SNORD26
chr3	12881810	12881949	NR_002582.1.1	38	-	SNORA7A
chr17	7478030	7478165	NR_002918.1.1	4	+	SNORA48
chr11	75111434	75111582	NR_000005.1.1	186	+	SNORD15A
chr20	47895481	47895560	NR_002433.1.1	11	+	SNORD12C
chr9	136217310	136217382	NR_002448.1.1	4	+	SNORD36A
chr1	76253573	76253657	NR_002749.1.1	5	+	SNORD45A
chr19	49994163	49994229	NR_000019.1.1	410	+	SNORD34
chr20	37062504	37062642	NR_003018.1.1	12	-	SNORA71D
chr11	62620381	62620507	NR_000008.1.1	2	-	SNORD22
chr6_ssto_hap7	2839682	2839760	NR_003065.1.1	6	-	SNORD84
chr1	109642814	109643234	NR_003023.1.1	8	+	SCARNA2
chr6_ssto_hap7	2834955	2835031	NR_003140.1.1	23	-	SNORD117
chr1	45241536	45241610	NR_000015.1.1	6	+	SNORD55
chr11	9450312	9450501	NR_002962.1.1	111	+	SNORA23
chr12	6619387	6619717	NR_004387.1.1	54	+	SCARNA10
chr14	20811229	20811570	NR_002312.1.1	5	-	RPPH1
chr3	129116052	129116191	NR_002992.1.1	46	-	SNORA7B
chr2	234197321	234197587	NR_003006.1.1	5	+	SCARNA6
chr19	17973396	17973529	NR_000012.1.1	6	+	SNORA68
chr2	86362992	86363129	NR_004378.1.1	7	+	SNORD94
chr4	119200344	119200475	NR_002963.1.1	218	+	SNORA24
chr18	47018033	47018099	NR_002572.1.1	880	-	SNORD58B
chr1	76252756	76252834	NR_003042.1.1	69	+	SNORD45C
chr17	7481272	7481409	NR_002912.1.1	308	+	SNORA67
chr20	37055948	37056086	NR_002911.1.1	289	-	SNORA71A
chr1	28906275	28906405	NR_002987.1.1	526	-	SNORA61
chr11	811680	811814	NR_002585.1.1	3317	+	SNORA52
chr12	7076499	7076769	NR_003010.1.1	89	-	SCARNA12
chr1	173833506	173833583	NR_002746.1.1	3479	-	SNORD47
chr16	2012334	2012467	NR_002327.1.1	12	-	SNORA10
chr14	21865451	21865560	NR_002916.1.1	13	-	SNORD8
chr11	75115464	75115610	NR_000025.1.1	14598	+	SNORD15B
chrUn_gl000220	155996	156152	NR_003285.1.1	2	+	RNA5-8S5
chr16	89627837	89627909	NR_002450.1.1	41	+	SNORD68
chr14	95999691	95999966	NR_003002.1.1	436	-	SCARNA13
chr1	567704	567793	NR_106781.1.1	8	-	MIR6723
chr18	47017652	47017717	NR_002571.1.1	479	-	SNORD58A
chr3	186505401	186505538	NR_002588.1.1	10	+	SNORA4
chr14	21860309	21860412	NR_003029.1.1	100	-	SNORD9
chr1	28906892	28907024	NR_002976.1.1	364	-	SNORA44
chr9	35657747	35658015	NR_003051.1.1	2948	-	RMRP
chr11	8706985	8707116	NR_002977.1.1	233	+	SNORA45B
chr20	2635712	2635844	NR_002981.1.1	13	+	SNORA51

Table 4.1. **List of excluded PTES transcripts from single exon genes.** PTES from single exon loci, including snoRNA genes, have high counts, are likely artefacts and not produced by backsplice.

Further analysis by Dr. Jackson (IGM, Newcastle University), comparing abundance of these snoRNA-PTES transcripts to that of linear snoRNAs in ENCODE small RNA samples (SRA ids: SRR446400-01) established that in all cases the PTES read counts were very low compared to canonical read counts in the small RNA fraction (see appendix 9.3). This investigation revealed five (from *SNORD34, SNORD15B, SNORD58B, SNORD47* and *SNORA52*) with the highest abundance ratios. These snoRNA-PTES were then chosen for further analyses that fully characterized their structure and origin of reads supporting them. Read density analysis of these transcripts suggest that supporting reads are likely lariat derived and do not emanate from *bona fide* PTES events. Figure 4.3A shows read densities of *SNORD34* on the UCSC genome browser, highlighting the difference in read distribution between cellular compartments.

First, there seems to be a lack of consensus in annotations for this gene. GENCODE v. 19 annotation reports this gene to be 71bp and extends beyond the RefSeq annotation (size: 66bp). PTESFinder evaluation of putative PTES transcript from this gene is annotation dependent and additional runs of PTESFinder using the GENCODE annotation did not identify this PTES transcript. Second, due to sequence identity between sequence at the 5' terminal and sequence flanking the 3' terminal, supporting reads map linearly (Fig 4.3B) and should have been filtered out. However, these reads escape filters for two reasons: 1) the host gene is a single exon gene and there is no canonical junction for mapping quality comparison, as the sequence region with high identity is intronic 2) the gap between the 3' terminal and the region of identity negates the genomic filter. Furthermore, read density pattern around *SNORD34* shows a consistent level of reads across the gene and surrounding intronic region (of *RPL13AP5* - a non-coding RNA), but there is an abrupt decrease towards the 5' splice site of *RPL13AP5* exon 6. The decrease in read density appears after the conserved adenosine residue, suggesting that reads within this region were derived from lariat intermediates. Additional analysis of sequence flanking *SNORD34* did not reveal inverted complementary regions that may aid circularization or trans-splicing.

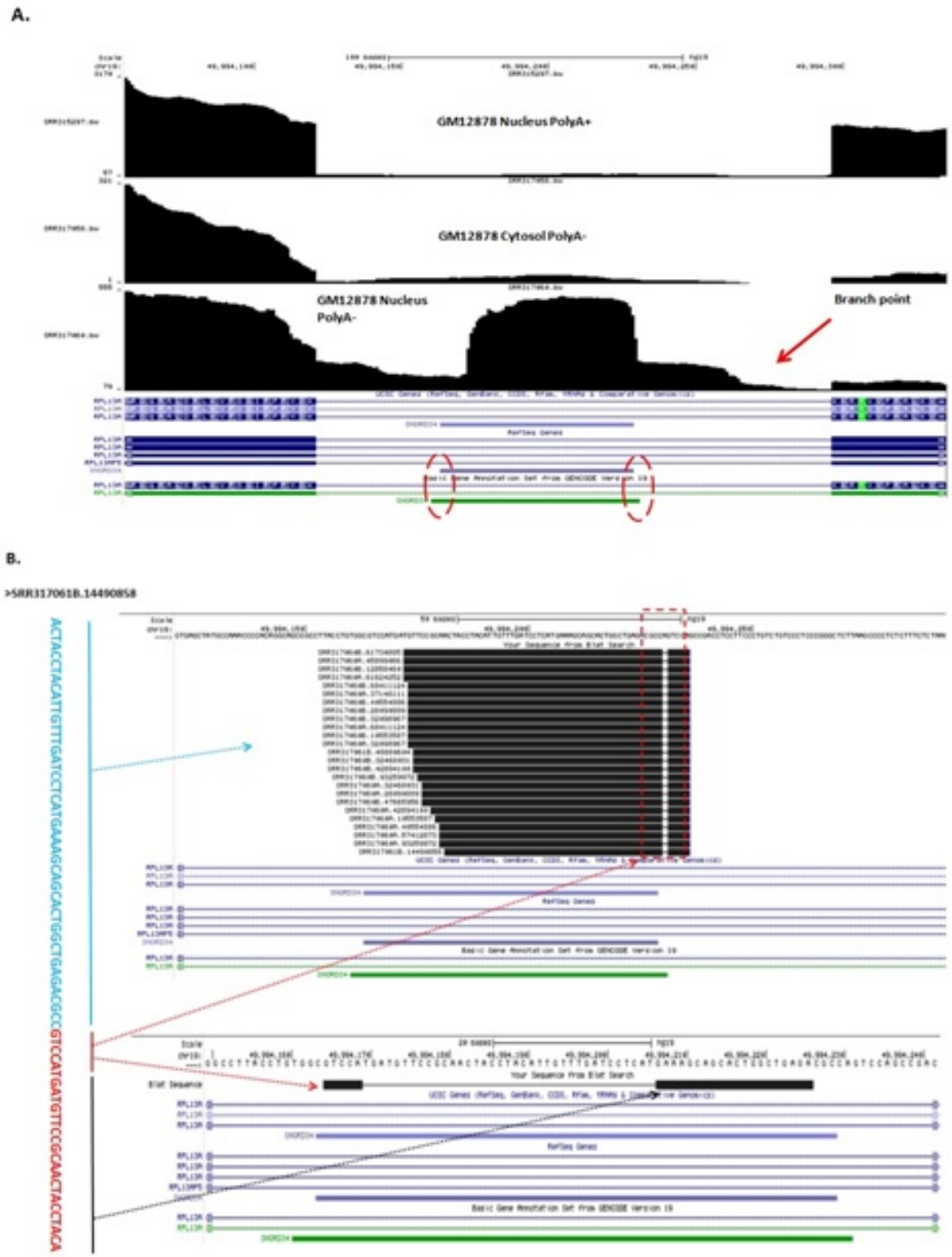


Figure 4.3. **Sequence analysis of SNORD34.1.1.** A) Read density analysis of *SNORD34* shows that reads from this genomic region are likely lariat derived and lack of concordance between GENCODE v 19 annotation of *SNORD34* and RefSeq's. B) Reads supporting putative PTES transcripts map linearly and extends to intronic sequence, due to sequence identity. Portions of reads also map to internal regions of the gene, in a pattern explainable by template switching or self-priming from 3' terminal.

Circularity of this putative PTES structures was investigated by RT-PCR using outward facing primers. RNA extracts from the nucleus and cytosol of HEK293 cells were treated with RNase R, an enzyme that degrades linear transcripts, enriching for circRNAs. Variability in band intensities for expected amplicons from *ANRIL* was observed (Fig 4.4). As RNase R removes linear molecules and enriches for circRNAs, increased band intensities were expectedly observed with higher concentrations (60U) of RNase R in both cellular compartments. However, various factors may explain the higher band intensities of *ANRIL* amplicons in the untreated samples, including: 1) amplification of both linear and circular RNA molecules with the same primer pairs, as multiple PTES events have been confirmed from this gene (Jeck & Sharpless, 2014; Burd et al., 2010); and 2) the reported sensitivity of some circRNAs to RNase R following linearization (Jeck et al., 2013). Although uneven gel loading during electrophoresis cannot be ruled out, it is clear that the previously confirmed PTES from *ANRIL* is resistant to RNase R. In contrast, the expected amplicon from *SNORD34* was not observed in samples treated with RNase R or in cytosolic fractions. However, the expected fragment size was observed in untreated sample from the nucleus (lane 5, Fig 4.4), suggesting that the structure is not circularized. The structure is also not enriched in GM12878 PolyA+ sample, as there are only 3 supporting reads, suggesting that it is not polyadenylated and most likely a linear reverse transcription artefact.

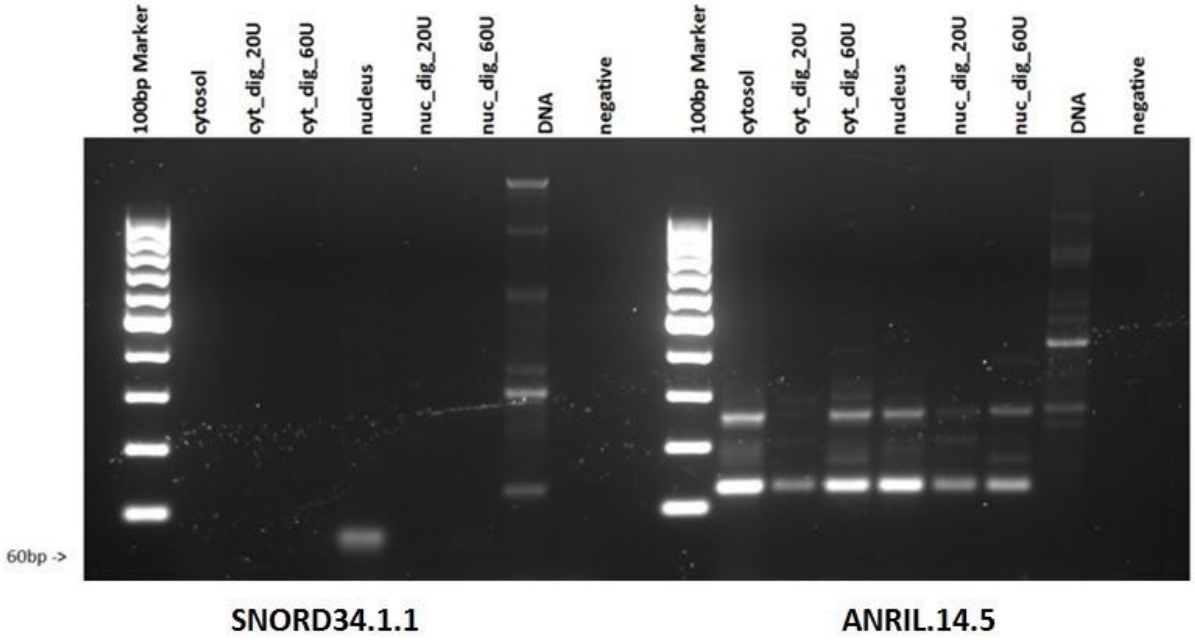


Figure 4.4. ***In vitro* confirmation of SNORD34.1.1**. Gel electrophoresis image showing expected amplicon size (60bp) for SNORD34.1.1 in nuclear RNA fraction, but not observed in cytosolic fractions and in samples treated with RNase R to remove linear molecules. Previously confirmed circRNA (ANRIL.14.5 [Burd et al., 2010]) is used as control.

Another example is the putative PTES transcript from *SNORD15B*. Reads supporting this structure were aligned to the genome (HG19) using BLAT (Kent 2002). Many of the reads map in a split manner, resembling *bona fide* PTES supporting reads. However, these reads perfectly map to the 3' region of *SNORD15B* and 5' region of *SNORD15A*, an upstream snoRNA of similar size (148bp), suggesting high sequence identity between both genes. Both snoRNAs are embedded within introns of Ribosomal protein S3 (*RPS3*). A further pairwise alignment between both snoRNA sequences confirmed the presence of homologous regions (Fig 4.5). Thus, reads reported as evidence for this PTES are most likely lariat derived or resulted from template switching artefacts.

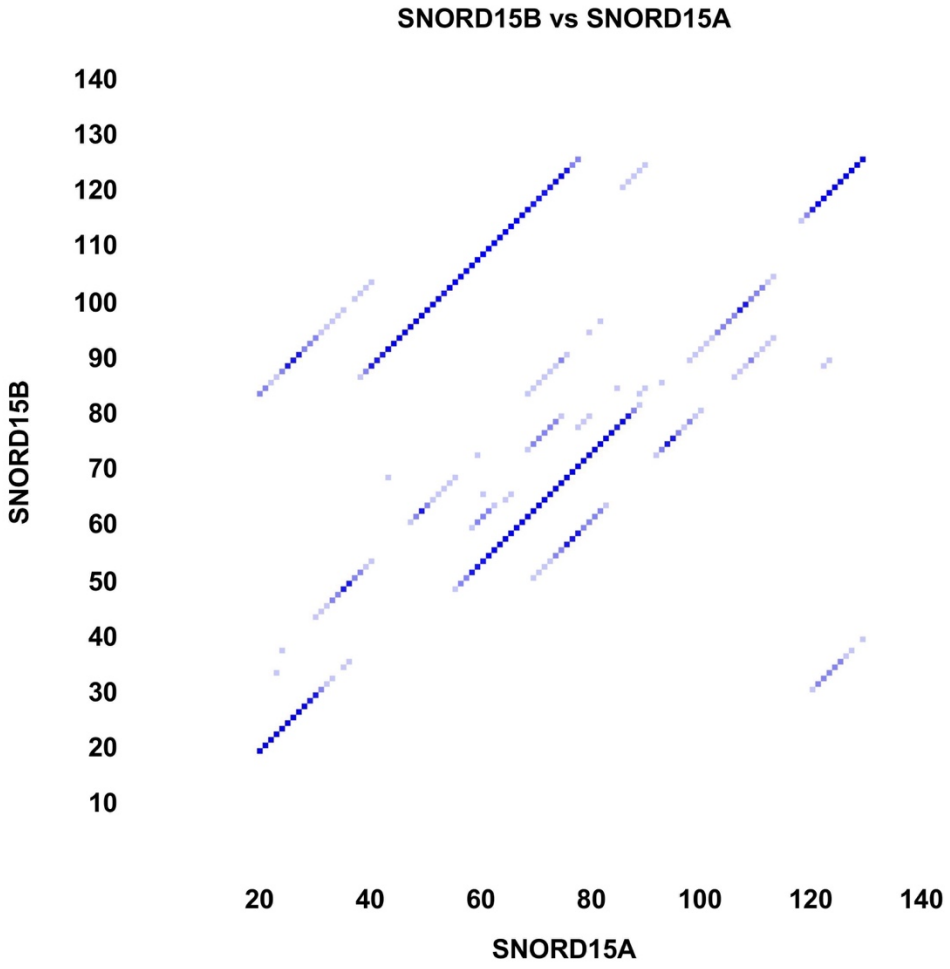


Figure 4.5. **Homologous regions of SNORD15A and SNORD15B.** Dot plot of pairwise alignment between SNORD15A and SNORD15B, showing regions of sequence homology that confound PTES discovery.

A final example is a PTES transcript from *RMRP*, a single exon endoribonuclease gene. Many of the reads supporting this putative PTES transcript are consistent with alignments expected for *bona fide* PTES events. To investigate the accuracy of this prediction further, a strand specific read density analysis of this gene was carried out. Unexpectedly, one peak each

was observed at opposite ends of the gene and on different strands (Fig 4.6). This pattern is only explainable by a combination of self-priming from one terminal and sense-antisense template switching. As these transcripts lack GT-AG splice sites and are mostly intronic, they do not conform to structures expected from PTESFinder runs. Additionally, there are no associated canonical junction counts for downstream comparative analysis. Because of the alternative explanations established for the most abundant structures analyzed, all 51 PTES from mono-exonic loci were excluded from enrichment analyses.

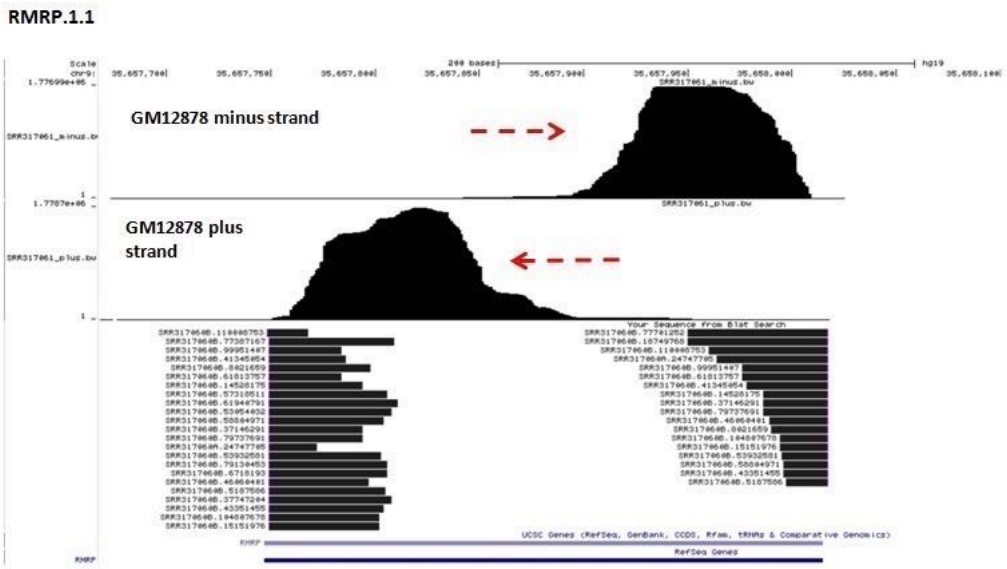


Figure 4.6. **Read density analysis for RMRP.1.1**. Reads supporting this putative PTES map in a split head-to-tail manner, akin to bona fide PTES reads. Read density analysis of both strands indicates that reads from this locus conceivably emanated from sense-antisense template switching artefacts.

4.3.3 PTES transcripts are enriched in the cytosol

The high overlap of PTES transcripts identified in both cellular compartments suggests that a pathway may exist for the export of these transcripts from the nucleus to the cytoplasm. It has been theorized that these transcripts may exit the nucleus during cell division (Jeck et al., 2013). Identifying transcripts retained (or specifically enriched) in the nucleus may aid our understanding of how these transcripts reach the cytoplasm. Enrichment in the cytoplasm may point to high stability of enriched transcripts and their functional significance. To assess the relative distributions of these transcripts within both compartments, I performed enrichment analyses using summed raw counts from all samples of each identified PTES transcript from each cellular compartment to reduce the effect of low counts in some samples (see 2.6.5 for details). From these tests, 2438 had a p-value less than 0.05. To correct for multiple testing, the

Benjamini-Hochberg protocol was used to control the false discovery rate at 0.05. This correction resulted in only 1048 reaching significance (Fig 4.7A).

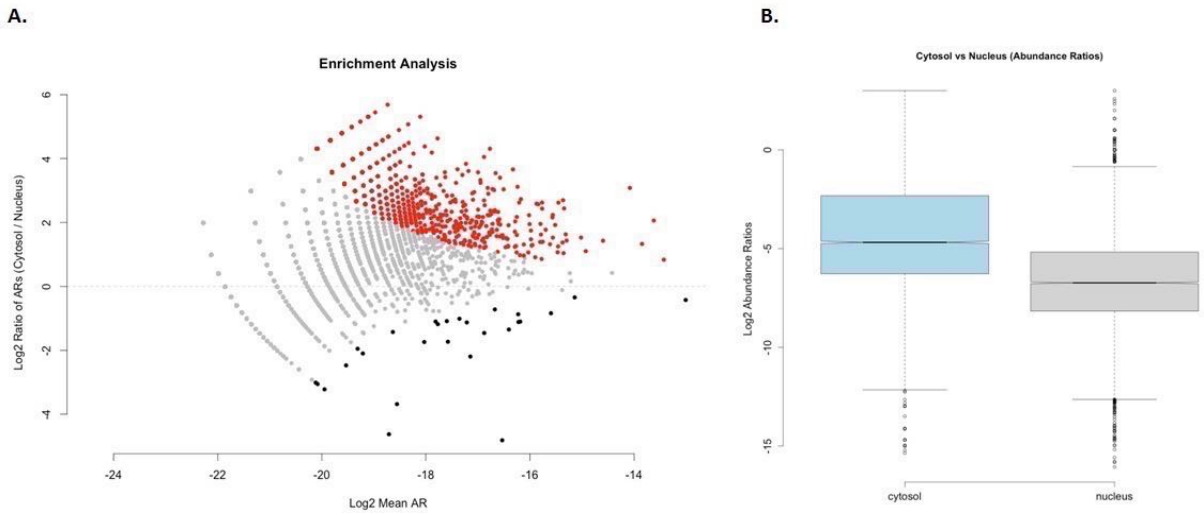


Figure 4.7. Sub-cellular PTES enrichment. A) Abundance ratios, derived by dividing PTES raw counts with total canonical junction counts (see methods), are plotted. Transcripts reaching significance after Fisher's exact tests are highlighted; enriched in cytosol (red); enriched in nucleus (black). B) Distributions of abundance ratios.

Of the 1048 identified as significant, 1007 (~9% of transcripts tested) were enriched in the cytosol, having an odds ratio greater than 1, when compared to abundance in the nucleus. Only 41 were enriched in the nucleus (Table 4.2), 13 of these appear to be nuclear-specific, with no reads observed in any cytosolic sample. Abundance ratios (AR) were derived for each PTES transcript by dividing PTES junction counts with total canonical junction counts observed in PTES producing gene. Box plots in Fig 4.7B show that PTES transcripts have higher abundance ratios in the cytosol (median: 0.04) than in the nucleus (median: 0.009). Interestingly, 114 transcripts - comprising of single exons - have AR > 1 in the cytosol and only 29 have AR > 1 in the nucleus.

chromosome	start	stop	structure	strand	host_gene	ptes_counts_cyt	cjuncs_counts_cyt	ptes_counts_nuc	cjuncs_counts_nuc	odds_ratios	pvalues	padj
chr22	40749076	40750331	NM_000026.4.3	+	ADSL	29	803	293	2483	0.322	2.50E-11	5.75E-09
chr2	97810254	97810449	NM_001164315.10.9	+	ANKRD36	4	91	447	34248	0.029	8.46E-46	3.37E-42
chr2	97858717	97867967	NM_001164315.47.38	+	ANKRD36	0	0	48	22500	0.000	3.69E-06	0.000220505
chr2	97860462	97870000	NM_001164315.50.39	+	ANKRD36	0	0	29	13140	0.000	0.000656337	0.014075306
chr2	98158373	98173595	NM_025190.28.11	-	ANKRD36B	0	0	308	5189	0.000	2.65E-36	7.91E-33
chr2	98154630	98171717	NM_025190.32.13	-	ANKRD36B	2	48	102	11224	0.064	7.65E-10	1.19E-07
chr2	98175428	98177200	NM_025190.9.8	-	ANKRD36B	0	0	38	11959	0.000	6.24E-05	0.002139672
chr12	111991961	111993723	NM_002973.3.2	-	ATXN2	11	400	145	3025	0.247	7.61E-08	7.76E-06
chr9	*138773478*	*138774924*	*NM_015447.3.2*	-	*CAMSAP1*	*416*	*518*	*2215*	*1569*	*0.612*	*3.23E-22*	*3.21E-19*
chr19	47865732	47865950	NM_014681.6.6	+	DHX34	39	315	283	1874	0.449	3.36E-07	2.91E-05
chr1	21415630	21437876	NM_001198803.5.4	-	EIF4G3	30	468	196	3424	0.498	0.000153835	0.004605404
chr1	21415630	21415706	NM_001198803.5.5	-	EIF4G3	0	0	24	1102	0.000	0.002750615	0.038544445
chr11	92085261	92088570	NM_001008781.1.1	+	FAT3	8	66	106	309	0.246	6.58E-06	0.000333016
chr17	79660683	79661883	NM_004712.12.10	+	HGS	17	826	136	6647	0.407	0.000124423	0.003772175
chr22	24056373	24057381	NR_024448.3.2	-	KB-1572G7.2	1	0	32	784	0.102	0.003117705	0.03928374
chr5	113740134	113740553	NM_021614.3.3	+	KCNN2	14	10	118	148	0.386	0.000280183	0.006706977
chr8	95549330	95550574	NM_015496.4.3	-	KIAA1429	37	985	315	3407	0.382	6.86E-10	1.11E-07
chr4	128995614	128996148	NM_001278604.3.2	+	LARP1B	3	55	51	1254	0.192	0.00102093	0.017725297
chr7	39818869	39822398	NR_026999.8.4	+	LINC00265	0	0	28	462	0.000	0.001109809	0.019101828
chr10	38734344	38736649	NR_024497.7.6	+	LINC00999	36	70	309	721	0.379	7.04E-10	1.12E-07
chr2	39559057	39564722	NM_001270425.8.5	-	MAP4K3	6	443	64	8040	0.305	0.001768882	0.028399596
chr19	9007486	9012898	NM_024690.43.34	-	MUC16	1	159	33	2368	0.099	0.001961049	0.030580593
chr15	59510089	59564648	NM_004998.10.2	-	MYO1E	12	642	108	6172	0.362	0.000218077	0.005827583
chr5	176618884	176639196	NM_172349.6.4	+	NSD1	0	0	28	1680	0.000	0.001109809	0.019101828
chr5	176636636	176639196	NM_172349.6.6	+	NSD1	0	0	22	1810	0.000	0.004385007	0.04997987
chr11	71671795	71671937	NR_024147.2.2	+	RNF121	0	0	25	587	0.000	0.001703625	0.02753694
chr2	55252221	55255356	NM_020532.3.2	-	RTN4	2	243	44	174	0.148	0.000771167	0.014978191
chr3	127806548	127806651	NM_003707.9.9	-	RUVBL1	1	66	37	1036	0.088	0.000821139	0.015845727
chrX	134988569	134988710	NM_018666.7.7	+	SAGE1	12	105	218	1167	0.179	1.07E-13	3.28E-11
chrX	134990247	134992329	NM_018666.15.11	+	SAGE1	0	0	32	1167	0.000	0.000257314	0.006301852
chr5	69515744	69517842	NR_024054.3.2	-	SMA4	97	645	488	8003	0.647	4.97E-05	0.001756876
chr15	67004005	67008836	NM_001142861.3.2	+	SMAD6	3	24	46	276	0.212	0.002119339	0.032749682
chr17	20107645	20109225	NM_001243439.4.4	+	SPECC1	61	231	433	1693	0.459	5.47E-10	9.07E-08
chr12	121220457	121222396	NM_139015.6.4	-	SPPL3	18	277	156	734	0.376	1.42E-05	0.000618664
chr19	1011578	1011662	NM_001033026.8.8	-	TMEM259	0	0	52	4106	0.000	1.45E-06	9.97E-05
chr12	*113705647*	*113707650*	*NM_017901.7.5*	+	*TPCN1*	*12*	*131*	*102*	*4158*	*0.383*	*0.000561013*	*0.012364019*
chr22	17117929	17119630	NR_001591.6.5	+	TPTEP1	20	106	218	333	0.299	2.16E-09	3.07E-07
chr5	82832825	82850857	NM_004385.12.8	+	VCAN	0	0	32	1695	0.000	0.000257314	0.006301852
chr5	37438023	37443474	NM_018034.7.6	+	WDR70	1	48	98	1178	0.033	2.06E-10	3.77E-08
chr16	28163979	28167848	NM_001270940.9.8	-	XPO6	0	0	29	13769	0.000	0.000656337	0.014075306
chr16	88552346	88555561	NM_153813.3.2	+	ZFPM1	0	0	22	306	0.000	0.004385007	0.04997987

Table 4.2. PTES transcripts significantly enriched in the nucleus. List of nuclear-enriched PTES transcripts, 13 were exclusively observed in nuclear fractions, with no detectable reads in cytosolic samples. Odds ratios were derived by dividing the abundance ratios (PTES counts / Canonical junction counts) of each transcript the cytosol by the abundance ratios in the nucleus.

4.3.4 Incompletely processed circRNAs observed in the nucleus

Many in the list of 41 transcripts significantly enriched in the nucleus are short relative to other identified transcripts; with median internal intron size of 884bp, compared to 1997bp for transcripts enriched in the cytosol. Within this list is a circRNA from the *CAMSAP1* locus (p-value: 3.21E-19), previously reported to have an intron containing isoform (Salzman et al., 2013; Zhang et al., 2014). Results obtained from the enrichment analysis suggest that the presence of this isoform within the nucleus (and absence in the cytosol) may be responsible for the difference in abundance between compartments. Read density analysis confirms the presence of this isoform in the nucleus of various cell lines, as intronic reads are observed with comparable read counts to reads mapped to the exons. Figure 4.8A-C shows read distributions across exons within *CAMSAP1*, with the highest number of reads observed between exons 2 and 3, including the intervening intron (Fig 4.8A-B). The read distribution across the intervening intron in the polyA- nucleus sample is different from the distribution observed in the cytosol and polyA+ samples, suggesting that the intron is only detectable in non-polyadenylated RNA and likely contained within circRNA. Analysis of flanking canonical junctions and a canonical junction between exon 1 and exon 4 suggests that the circularization of CAMSAP1.3.2 may occur in the absence of alternative splicing, as read counts for these junctions were relatively low or absent in most samples. For instance, in GM12878 nucleus, a total of 221 reads map to CAMSAP1.3.2 PTES junction, only 28 reads map to both flanking canonical junctions (between exon 1 - exon 2 and exon 3 - exon 4). A single read supports the splice between exons 1 & 4 in both biological replicates. Multiple complementary inverted repeats were identified within flanking introns, suggesting that the backsplice between exon 3 & 2 is mediated by intron pairing.

Another example of this phenomenon may be the PTES transcript - comprising of exons 6, 7 and 8 of *TPCN1* locus (Fig 4.9) - enriched in the nucleus (p-value: 0.01). Similar to the structure from *CAMSAP1*, no reads support expected canonical splice junctions suggestive of exon skipping, indicating that this transcript may not be lariat derived or result from alternative splicing. From the read density pattern in GM12878 nucleus, two introns are presumably retained. Exons 7 and 8 are deplete of reads in the cytosol. This isoform is not observed in the cytosol and may contribute to ~3X enrichment of this PTES junction in the nucleus.

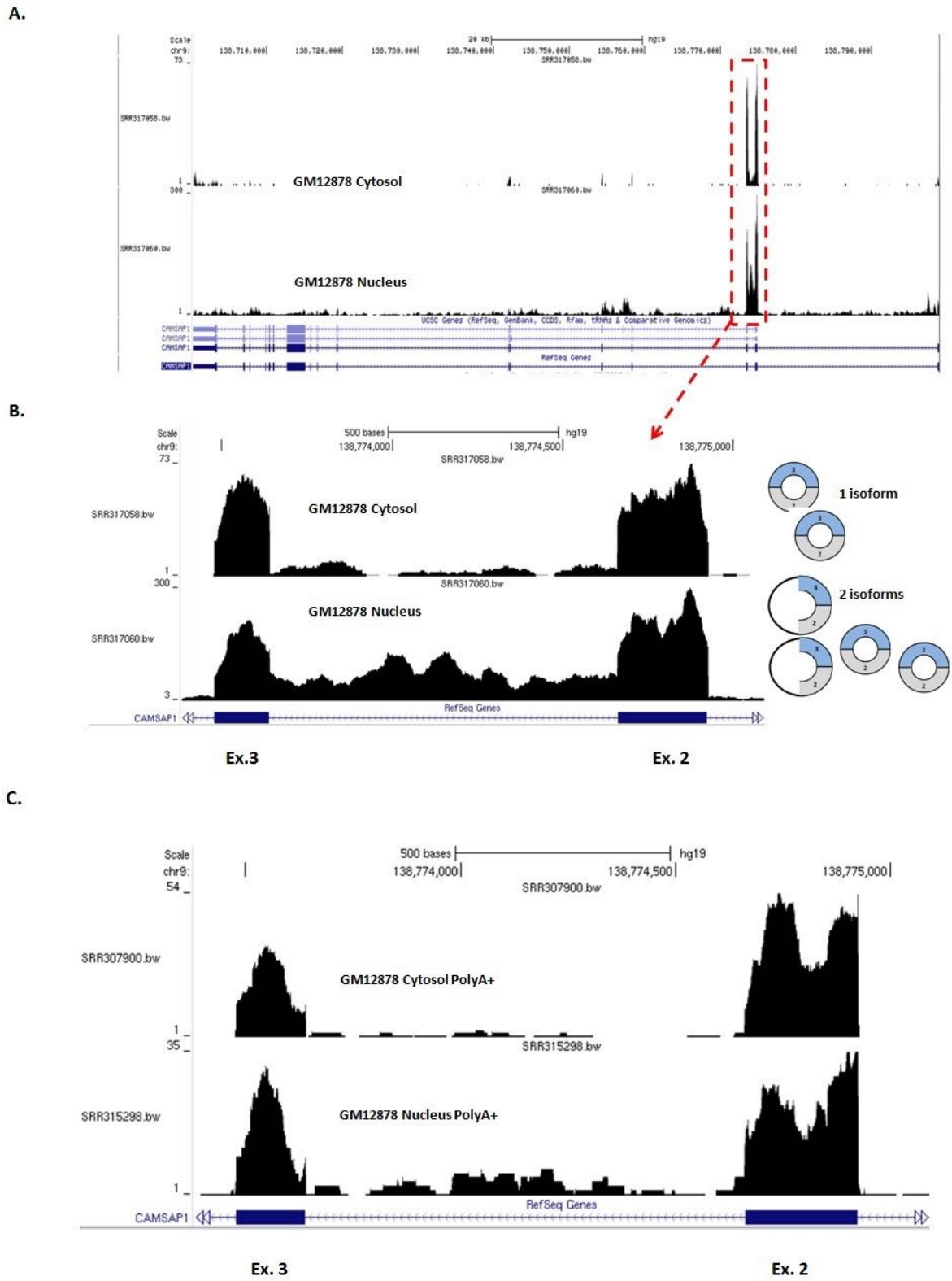


Figure 4.8. ***CAMSAP1* intron-retained PTES transcripts are enriched in the nucleus.** A) Read density of *CAMSAP1* locus, show highest read peak around exons 2 & 3. B) Two isoforms or circRNA involving exons 2 & 3 are observable in GM12878 nuclear fraction and only one isoform -without retained intron - is observed in cytosolic fraction. C) Read distribution pattern across exons 2 & 3 of *CAMSAP1* in PolyA+ Nucleus and Cytosolic fractions of GM12878.

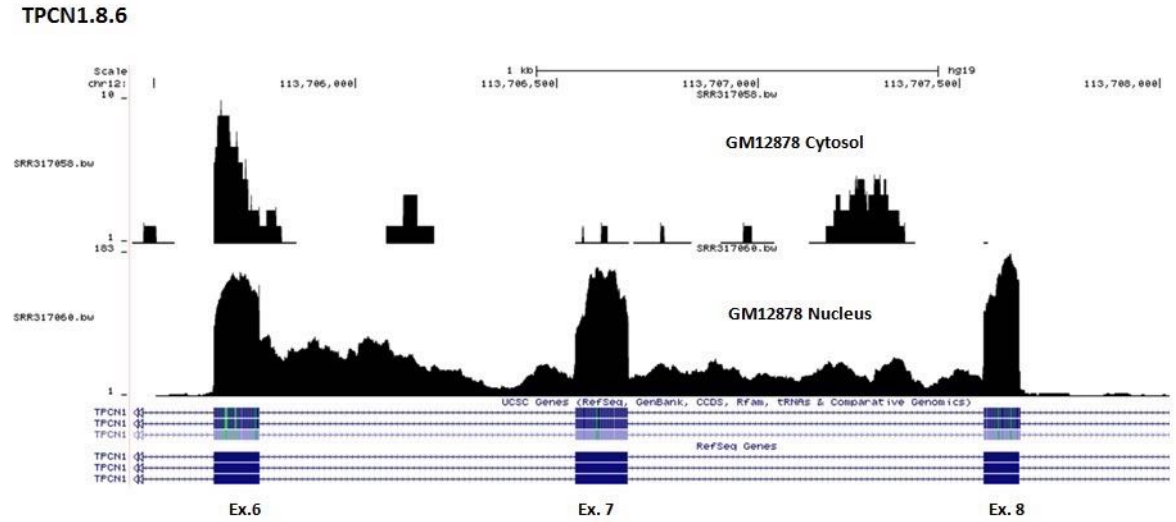
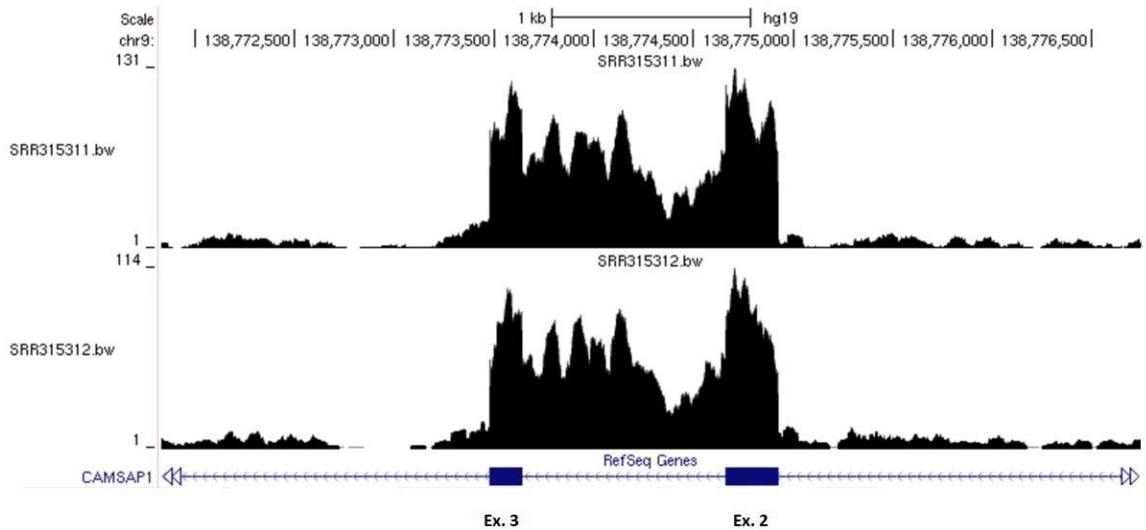


Figure 4.9. **Read distribution across exons of *TPCN1*.** TPCN1.8.6 is enriched in the nucleus, relative to the cytosol and exhibits read distribution patterns suggestive of intron retention in the nucleus.

4.3.5 Quantitative analysis of chromatin-associated PTES transcripts

Manual examination of read densities of CAMSAP1.3.2 and TPCN1.8.6 in other cellular compartments within the nucleus revealed that, at least in the case of CAMSAP1.3.2, intron-containing isoforms are present in the nucleolus (Fig 4.10). Intriguingly, read counts supporting CAMSAP1.3.2 are highest in the nucleolus, relative to samples of equivalent library sizes in other compartments within the nucleus (Fig 4.10 inset table). These transcripts are likely translocated away from the site of transcription in the nucleoplasm and into the nucleolus. This observation raises questions about whether PTES biogenesis occurs co-transcriptionally or after transcription.



PTES	Cytosol (n = 10)	Nucleus (n = 13)	Nucleoplasm (n = 2)	Nucleolus (n = 2)	Chromatin (n = 2)
CAMSAP1.3.2	416	2215	66	119	72
TPCN1.8.6	12	102	0	0	0

Figure 4.10. **Intron-retained CAMSAP1.3.2 in nucleolus.** Read distribution across exons of CAMSAP1.3.2 in 2 K562 nucleolus samples, suggests the presence of intron-retained PTES in nucleolus. Inset) Table showing number of supporting read counts for CAMSAP1.3.2 and TPCN1.8.6 in various cellular compartments; number of samples analyzed for each compartment shown in brackets.

From the analysis of chromatin-associated RNA samples, a total of 1246 distinct PTES transcripts were identified, the lowest number of transcripts identified from any cellular compartment. When compared with transcripts identified in the nucleoplasm, there is only about 35% overlap, highlighting possible sampling bias. Poor sequence quality of reads can impact PTES identification. To assess this, I examined the per base sequence quality of both biological replicates using FASTQC (Fig. 4.11). Results show that, for one sample, the last 20bps of reads have poor quality scores and possibly contribute to the number of PTES transcripts identified. Trimming the last 20bp of each read in both samples and reanalyzing with PTESFinder, identified 771 additional PTES, highlighting the impact of poor sequence quality on PTES discovery.

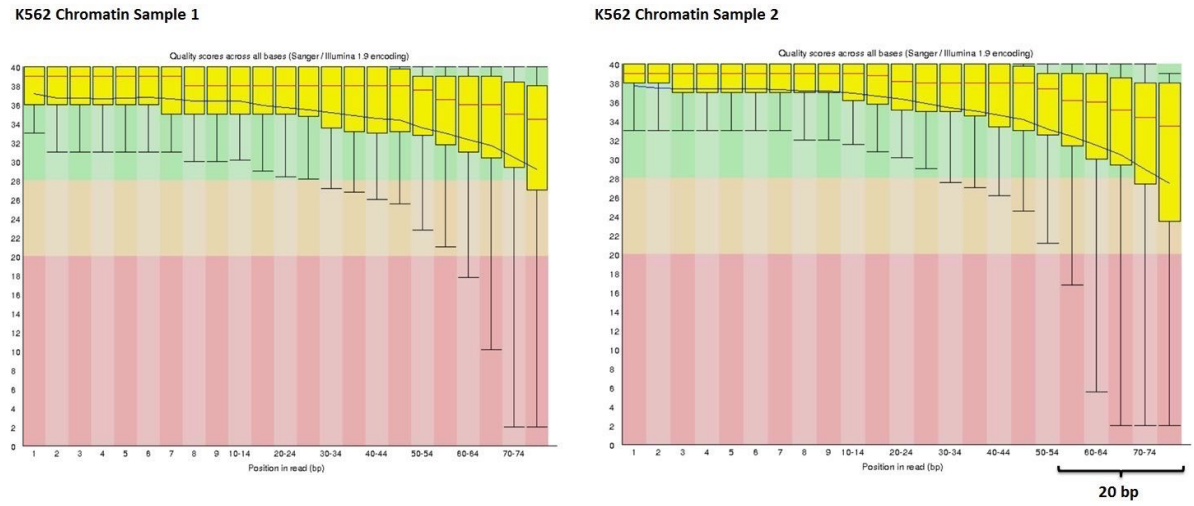


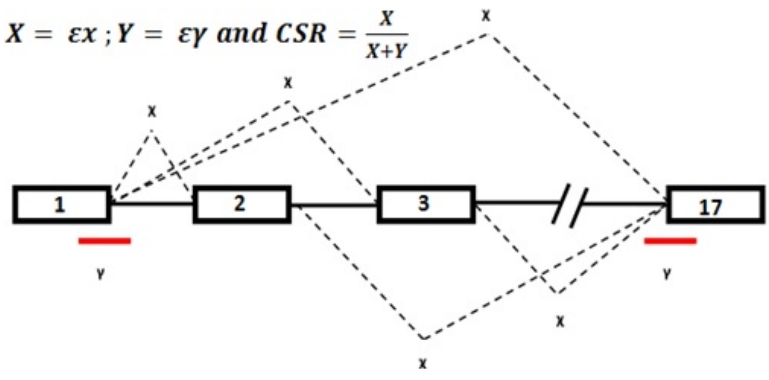
Figure 4.11. **Read quality of K562 chromatin RNAseq samples.** Results of FASTQC analysis of reads in chromatin samples. Nucleotide positions (highlighted - right) with box and whiskers extending into the red region have low confidence quality scores.

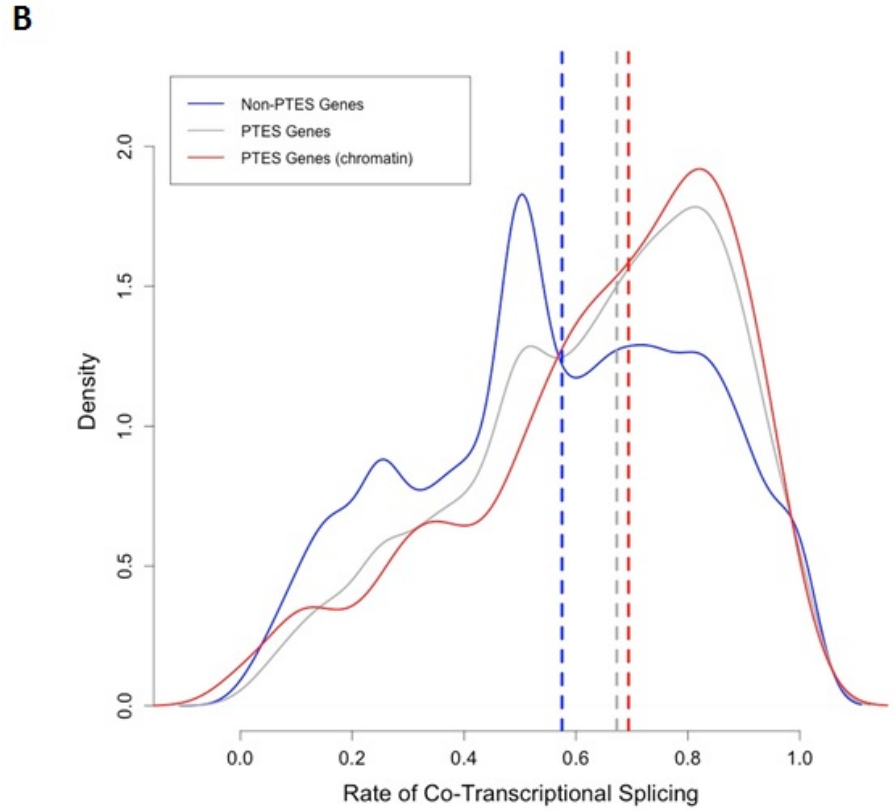
As some PTES may associate with chromatin, facilitating transcription of linear transcripts from their loci (Li et al., 2015) and RNA fractions can be contaminated during fractionation, identifying PTES within chromatin samples may not conclusively indicate that PTES occurs during transcription. For this reason, I assessed the level of co-transcriptional splicing for both PTES producing genes and genes without an identified PTES transcript in all compartments. Using *in silico* methods, I derived the co-transcriptional splicing rates (CSR) for all genes, as depicted in Figure 4.12A. My method (see 2.6.4 for details) is similar to the method described in Tilgner et al., (2012), where only a subset of internal exons was analyzed. Both methods compare expression of pre-mRNAs to that of mature (spliced) mRNAs of the same gene.

The CSR profiles for PTES producing genes (excluding those found in chromatin), PTES producing genes (chromatin-associated transcripts only) and non-PTES producing genes are presented in Fig. 4.12B; medians of 0.563, 0.667 and 0.686 respectively. PTES producing genes seemingly undergo more co-transcriptional splicing than non-PTES producing genes. Interestingly, host genes of PTES transcripts found in chromatin undergo more co-transcriptional splicing than other PTES producing genes. As ncRNAs are spliced more inefficiently than protein-coding genes (Tilgner et al., 2012), it may be that the CSR patterns observed were due to the number of ncRNAs in each gene set. To investigate whether this effect is observed, I removed all ncRNAs from each gene set and repeated the analysis. This resulted in median CSRs of 0.575, 0.673 and 0.694 respectively, a marginal increase, but the patterns remained.

Comparing the CSR of terminal exons (1 & 17) of *CAMSAP1*, it is striking that the first exon is almost always spliced (CSR = 1), unlike the last exon (CSR ~ 0.33). Reads mapping to the exon-intron junction of the last exon are easily detectable suggesting that the backsplice of exon 3 and exon 2 may occur co-transcriptionally and released from the 'fractured' pre-mRNA. With mean CSR of 0.96, *TPCN1* appears to be spliced fully co-transcriptionally. It is conceivable that the fraction of transcripts not completely processed co-transcriptionally result in PTES transcripts. However, the absence of reads mapped to exon-intron junctions of both *CAMSAP1* & *TPCN1*, suggests that PTES in these loci (and perhaps many others) occur co-transcriptionally. Notably, the vast majority of PTES producing genes have their first introns removed co-transcriptionally (Fig 4.12C). Consistent with CSRs of PTES producing genes derived for the whole gene, their first and last exon CSRs are higher than that of non-PTES genes. However, taking the ratios of first and last exon CSRs for both groups, the reduction in CSR is highest for PTES genes (median: 1.47), relative to that of non-PTES genes (median: 1.17). A non-parametric test suggests this difference to be significant (p-value: 2.2e-16, Wilcoxon rank sum test).

A





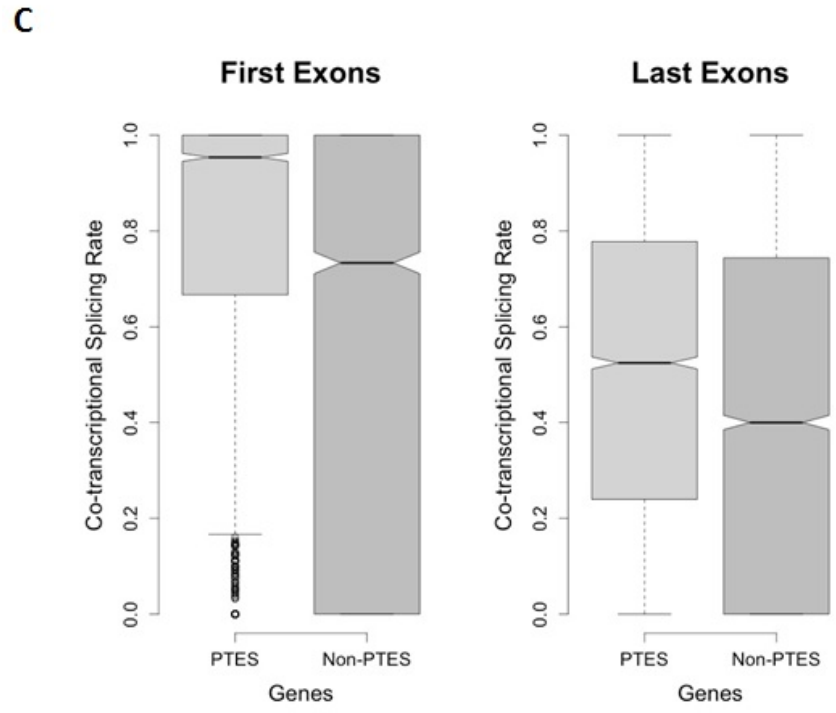


Figure 4.12. **Co-transcriptional PTES biogenesis**. A) Co-transcriptional splicing rates (CSR) were derived by summing the number of reads mapped to splice junctions involving terminal exons and dividing by the number of reads mapped to their exon-intron junctions. *CAMSAP1* gene model is used for illustration. B) CSR profiles of all genes without an identified PTES transcript (blue), all PTES genes, excluding host genes of transcripts identified in chromatin (grey) and PTES producing genes of transcripts identified in chromatin. Median CSRs are 0.563, 0.667 and 0.686 respectively - highlighted with vertical lines. C) Distribution of CSRs, derived using first and last exons separately.

4.3.6 Do PTES transcripts contribute to the proteome?

To assess the protein-coding potential of PTES transcripts, RNAseq data from sucrose gradient fractionated HEK293 cells, treated with arsenite to induce translational arrest were obtained (see Karginov & Hannon 2013 for details). PTESFinder v. 1 was used to screen for PTES events in 16 samples, 8 untreated and 8 arsenite treated. Table 4.3 summarises the number of identified structures from each fraction in both conditions.

	Fractions	Untreated Samples			Arsenite Treated Samples		
		PTES Transcripts	PTES Reads	Reads / PTES ratio	PTES Transcripts	PTES Reads	Reads / PTES ratio
Light	1	2243	10257	4.57	1365	4615	3.38
	2	577	1250	2.17	357	782	2.19
	3	131	188	1.44	57	74	1.30
	4	158	259	1.64	105	167	1.59
Heavy	5	70	122	1.74	115	152	1.32
	6	49	88	1.80	131	198	1.51
	7	57	83	1.46	139	208	1.50
	8	73	141	1.93	151	280	1.85

Table 4.3. **PTES identified from HEK293 sucrose gradient fractions.** Summary table showing number of PTES and PTES supporting reads, identified from sucrose gradient fractions of HEK293, with and without arsenite treatment to inhibit translation.

Most of the structures identified were from the lighter fractions (fractions 1 - 4) and not polysome associated. To estimate abundance, reads were mapped to sequence references for ERCC spike-ins - a standard set of exogenous synthetic RNAs used as controls in gene expression analysis (Jiang et al., 2011; ERCC, 2005) - to estimate library size factors. Figure 4.13A shows sample size normalized raw counts for identified PTES structures in both arsenite treated and untreated samples. Results suggest little or no effect of arsenite treatment on number of reads supporting PTES events. Because, arsenite treatment inhibits translation initiation and elongation, translational activity is expected to be limited in fractions 4 - 8 upon treatment. In response to stress (arsenite treatment) the polysome is dislodged causing the accumulation of transcripts in the lighter fractions (Karginov & Hannon 2013). For comparisons, I extracted reads mapping to the last 120bp of last exons of all genes. As PTES transcripts contribute to canonical junction counts and terminal exons are rarely involved in PTES, raw counts from these exons should provide an unbiased estimate of canonical transcripts. Raw reads counts were then normalized using the same library size factors derived from ERCC spike-in expression levels. The observed pattern (Fig 4.13B) is noticeably different from the observed pattern for PTES transcripts. Furthermore, most genes undergoing translation are expected to

have abundance pattern similar to that of *ACTB* (Fig 4.13D). Although some genes are up-regulated or continue to be expressed upon stress (Karginov & Hannon 2013), most canonical transcripts appear to exhibit the same pattern as *ACTB*.

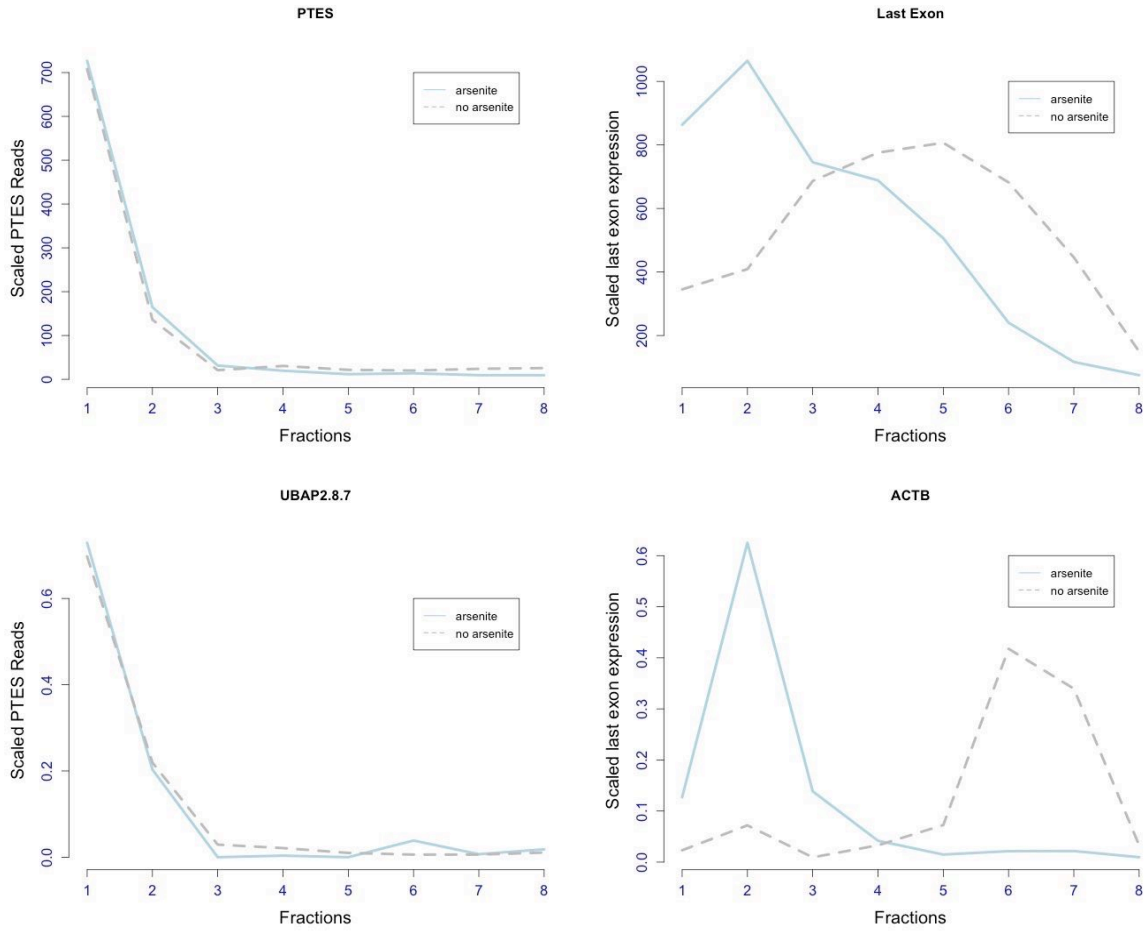


Figure 4.13. **Assessing translational potential of PTES transcripts**. Data from sucrose gradient fractionated samples (1 - 8), both arsenite treated (blue solid line) and untreated (grey dashed line), were analyzed with PTESFinder and are plotted. A) Normalized total read counts for discovered PTES structures. B) Normalized abundance pattern for last exons of all genes. C) Normalized abundance for PTES transcript from *UBAP2* locus D) Abundance of *ACTB* canonical transcript.

A structure from *UBAP2* (consisting of exons 8 and 7) is observed in 15 of the 16 samples. Previous studies (Jeck et al., 2013; Salzman et al., 2013; Rybak-Wolf et al., 2015) have identified this transcript from RNase R enriched samples, suggesting its circularity. At the sequence level, this PTES also has 4 AUG codons, with reads supporting PTES either starting at or mapping across 1 of the start codons near the end of exon 8. Further sequence analysis revealed 2 stop codons upstream of this start codon, indicating that an ORF is present and this transcript may be translated (appendix 9.3). Despite this, the pattern observed from this analysis suggests that arsenite treatment has no effect on this transcript (Fig 4.13C), that the vast majority of PTES transcripts do not associate with polysomes and are unlikely to contribute to the proteome.

4.4 Discussion

A generic RNA is transcribed, processed and translocated to its site of functional relevance. Various RNA species are processed differently and exported using different pathways. For mRNAs, it is increasingly evident that co-transcriptional processing occurs; mRNAs are made export competent by phosphorylation of serine-arginine proteins and other export adapter proteins, prior to nucleo-cytoplasmic export and translation. Cap structures, exon junction complexes and polyA tails aid in export of mRNAs to the cytosol, but many PTES transcripts lack these features. Although PTES transcripts have been identified from cytosolic RNA fractions, it remained unclear 1) whether their biogenesis occurs co-transcriptionally, 2) whether they are exported via known export pathways, 3) whether they contribute to the proteome once in the cytoplasm and 4) their relative distributions in sub-cellular compartments.

It had been speculated that perhaps, PTES transcripts exit the nucleus during mitosis (Jeck et al. 2013). My analysis of samples from sub-cellular compartments of various human cell lines revealed that, consistent with all transcription occurring in the nucleus, more PTES transcripts are detectable in the nucleus than the cytosol. Some PTES transcripts (with AR > 1) were observed to have abundance higher than observed for total canonical junctions from host genes. These transcripts do not contribute to the total canonical junctions of their respective loci. Previous estimates of PTES abundance vary. Nigro et al., (1991) estimated the abundance of PTES in the DCC gene to ~0.1% of whole gene expression. There are estimates of between 5 and 10% for transcripts reported in Salzman et al., (2013). The range of abundance ratios (AR) obtained in this analysis is consistent with previous reports, including for transcripts with expression values higher or comparable to linear counterparts described in Capel et al, (1993) and Al-Balool et al., (2011).

Enrichment analysis of PTES transcripts in both compartments revealed that ~9% of transcripts tested are significantly enriched in the cytosol relative to the nucleus. This enrichment bias may be due to accumulation of circRNAs in the cytosol, as a result of their resistance to exonuclease activity. In contrast, a smaller proportion of transcripts tested were significantly enriched in the nucleus. Majority of these nuclear-enriched transcripts are single-exon transcripts or have relatively short internal introns (median: 884bp, compared to 1997bp), as exemplified by CAMSAP1.3.2, which has a 1021bp intron between exons involved in PTES. An isoform of TPCN1.8.6 appears to have two short retained introns of sizes 785 & 884bp. Such introns likely elude removal due to weak splice signals and are retained within PTES transcripts. Read density analysis of CAMSAP1.3.2 and TPCN1.8.6 confirmed the presence of retained introns enveloped by backspliced exons. Two previous reports have confirmed the identification of an intron containing isoform of CAMSAP1.3.2 using *in vitro* methods

(Salzman et al., 2013; Zhang et al., 2014). The abundance of this isoform is reported to be tissue-specific (Salzman et al., 2013), perhaps dependent on variation in transcription elongation rates. It is also thought to be unstable, having a half-life of ~7 mins (Zhang et al., 2014), suggesting that the retained intron is removed quickly or that linearisation and subsequent degradation occurs. The differential expression of some single exon transcripts is perhaps explainable by miRNA induced endonucleolytic decay, once in the cytosol. Some of these transcripts harbour numerous miRNA binding sites; SAGE1.7.7 for instance, has at least 7 predicted miRNA binding sites within its 141bp sequence. MiRNAs have been shown to antagonise circRNAs, as exemplified by mir-671 activity on circCDR1 (Hansen et al., 2011). Moreover, although evidence of circCDR1 circularity has been provided (Hansen et al. 2011; Hansen et al. 2013; Memczak et al. 2013), Jeck et al., (2013) found a subset of presumed circRNAs (including circCDR1) not to be enriched in RNase R digested fibroblasts sample, relative to undigested samples. This is explainable by the presence of miRNA binding sites in PTES transcripts, resulting in their linearization. The observation of nuclear-enriched or nuclear-retained transcripts, nevertheless, support the presence of an unknown export pathway.

It is not currently clear if PTES biogenesis occurs co-transcriptionally or after release from chromatin. Conflicting pieces of evidence support PTES biogenesis during and after transcription, and the two are not necessarily mutually exclusive. Screening for PTES transcripts in chromatin-associated RNAseq data, a relatively small number of transcripts were identified, compared to transcripts identified from other compartments. This is likely due to poor sequence quality of the libraries analyzed or quick release from the chromatin post-transcription (Ashwal-Fluss et al. 2014). Co-transcriptional splicing analysis was performed by comparing the number of reads mapped to pre-mRNA exon-intron junctions to read counts from spliced junctions. Results revealed that PTES producing genes likely undergo more co-transcriptional splicing than non-PTES producing genes, suggesting that PTES biogenesis may indeed occur during transcription. This premise is consistent with reports indicating that PTES biogenesis does not require the formation of polyA tails (Ashwal-Fluss et al. 2014; Kramer et al., 2015). However, there is a significant difference between the rate of first intron removal and that of last introns, when PTES producing and non-PTES producing genes are compared. There is evidence that some transcripts with processing defects, including failure of snRNP to bind nascent transcript (Eberle et al. 2010), are retained at the gene (not released from chromatin) and subsequently targeted by the exosome (Bentley, 2014; Eberle et al., 2010). It is conceivable that PTES can induce such defects in nascent transcripts, resulting in their retention and lower comparable CSRs (depicted in Fig. 4.14). Such phenomena may impact the nuclear-cytoplasmic levels and subsequent expression of cognate linear transcripts. In the intron-pairing

biogenesis model, consider a hypothetical five exon PTES producing gene, the formation of secondary structure between intron 1 and intron 3, results in circularization between exons 3 and 2. As transcription proceeds, a splice between exons 4 and 5 may occur, but it is likely that intron 3 is not adequately removed from exon 4, inducing retention and subsequent degradation. This hypothetical case is consistent with suggested competition with linear canonical splicing (Ashwal-Fluss et al. 2014). In the re-splicing model, skipped exons are involved in PTES and the nascent transcript is unlikely to be impacted. The identification of an intron-containing PTES transcript from *CAMSAP1* in the nucleolus, away from site of transcription in the nucleoplasm, suggests that some PTES transcripts are incompletely processed prior to release from the chromatin.

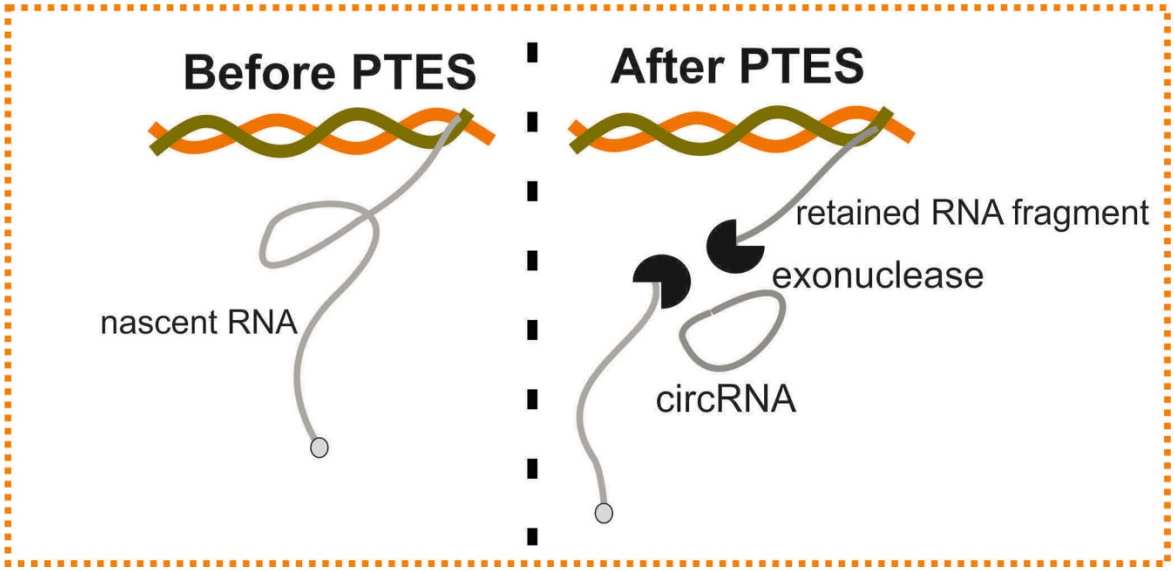


Figure 4.14. **Schematic illustration of impact of co-transcriptional circRNA biogenesis.** Co-transcriptional circRNA biogenesis is likely to induce splicing defects caused by failure to assemble spliceosomal proteins on the nascent transcript, subsequently resulting in retention at unclear speckles and degradation. Such an occurrence will limit the abundance of mRNAs exported to the cytosol.

Finally, analysis of sucrose gradient fractionated HEK293 cells, with and without arsenite treatment revealed that PTES transcripts do not contribute to the proteome. Arsenite treatment induces cellular stress and translational arrest by dislodging polysomes from translating RNA templates. These transcripts are subsequently observable in lighter fractions post-fractionation. Majority of PTES transcripts were identified from fractions containing transcripts with dislodged ribosomes, regardless of treatment group. The expression profiles of PTES transcripts identified from both treatment groups indicate that arsenite treatment has little or no effect. These transcripts are most likely not bound by polysomes, thus, do not contribute to the proteome. As experiments have shown, it is possible to derive protein products from circRNAs, particularly in the presence of ribosome entry sites and open reading frames (Wang & Wang 2014). Nevertheless, my results show that globally, PTES transcripts do not contribute to the

proteome. This conclusion is also supported by a recently published report by (Guo et al. 2014), where they found no PTES supporting reads within ribosome profiling data from a human bone osteosarcoma (U2OS) cell line.

4.5 Conclusion

In this chapter, I applied *in silico* methods to assess the distribution of PTES transcripts in various cellular compartments of 7 human cell lines. Expectedly, a variety of PTES transcripts were identified from nuclear fractions, including transcripts from single exon genes that are likely reverse transcriptase artefacts. Enrichment analysis of PTES transcripts found in both nucleus and cytosol, established that ~9% of PTES transcripts tested are enriched in the cytosol and < 0.5% (41) are enriched in the nucleus. Among the transcripts enriched in the nucleus are transcripts with intron-retained isoforms. These isoforms were also found in the nucleolus, suggesting incomplete processing prior to release from chromatin. PTES producing genes were however estimated to undergo more co-transcriptional splicing than non-PTES producing genes, suggesting that PTES biogenesis occurs during transcription for most genes. Enrichment of PTES transcripts in the cytosol may suggest possible translation once exported, however, results presented in this chapter show that, although some PTES transcripts may be bound by polysomes, arsenite treatment had no effect on ribosome occupancy, suggesting that they are not translated. In subsequent chapters, I assess the distribution of PTES in various human tissues, anucleate cells and during development, in an attempt to further elucidate the global properties of these transcripts and their functional significance.

Chapter 5. PTES transcripts in anucleate cells

5.1 Introduction

In the last chapter, a variety of PTES transcripts, including two with apparent retained introns, were identified from the nuclei of 7 human cell lines; and ~9% of identified PTES transcripts were shown to be significantly more abundant in the cytosol, relative to the nucleus. This enrichment in the cytosol is likely due to increased stability of circRNAs, conferred by lack of free termini, thus resistant to exonuclease activity. Enrichment in the cytosol does not however equate to contribution of PTES to the proteome, as was also shown in the last chapter. In this chapter, I extend my investigation of PTES distribution in cellular space to anucleate cells. Several reports have identified some circRNAs from cell-free RNA (Bahn et al. 2015; Lasda & Parker 2016), indicating that due to their physical properties, PTES transcripts remain detectable in the absence of steady transcription and efflux from the nucleus. However, PTES has not been reported from anucleate cells and no quantitative comparisons with cognate linear transcripts have been performed.

5.1.1 Platelets have complex transcriptomes

Platelets and mature erythrocytes are anucleate cells and lack genomic DNA. Platelets are derived from megakaryocytes following endoreduplication (DNA replication without cytokinesis), cytoplasmic expansion and release of cytoplasmic fragments (proplatelets) (Deutsch & Tomer 2006). There is evidence that platelet biogenesis results in rupture and apoptosis of the megakaryocytes progenitor cells (Nishimura et al. 2015). Typically, platelets are short-lived (~10 days), have roles in hemostasis and wound healing, becoming activated and aggregating around wounds to slow bleeding. Similar to exosomes, which are thought to have roles in cell-to-cell communication by transferring their cargo (proteins, lipids and RNA) between cells (Camussi et al. 2010), platelets have been shown to release micro-particles when activated (Risitano et al. 2012). These microparticles can contain RNAs that are transferred to other cells. Risitano et al., (2012) demonstrated that labelled RNA molecules (including GFP) in cultured megakaryocytes, were detectable in isolated platelets and were transferred to monocytes and HUVEC cells upon co-incubation with platelets.

Although platelets are understood to derive their transcriptomes from megakaryocytes progenitor cells and contain over 10,000-fold less mRNA than nucleated cells (Geiger et al. 2013; Landry et al. 2009), some transcripts within platelets uniquely undergo cytoplasmic splicing. Denis et al., (2005) observed the presence of spliceosomal proteins in circulating

platelets, suggesting the presence of the major spliceosome. The authors also demonstrated that resting platelets contain both pre-mRNA transcripts and trace amounts of mature mRNAs of interleukin-1B (*IL-1B*), a cytokine. Upon platelet activation, *IL-1B* pre-mRNAs are spliced, prior to translation in platelets (Denis et al. 2005). Similarly, miRNA biogenesis is thought to occur in platelets. Landry et al., (2009) identified pri-miRNAs and seemingly incompletely cleaved miRNAs of ~32 - 34bp in platelets. They further identified the complex of *Dicer* and *TRBP-2*, involved in miRNA maturation, suggesting that miRNA cleaving occurs in platelets (Landry et al., 2009). Taken together, it is therefore conceivable that novel PTES events may be identified from platelets.

5.1.2 Platelets transcriptomes vary between human donors

Various studies have identified less than 50% of annotated protein coding transcripts in platelets, consistent with reduced mRNA levels in these specialized cells (Bray et al. 2013; Best et al. 2015; Londin et al. 2014). Bioinformatics analysis of platelets transcriptomes from 4 human donors identified variation in number of detectable transcripts, ranging from 5511 to 10862 protein-coding genes. (Bray et al., 2013). Additionally, novel intronic transcripts, pervasive antisense transcripts and a subset of unmapped reads with no defined origins were identified from all donors, underscoring the diversity of transcripts within platelets (Bray et al., 2013). In another study, 10 donors from 2 ethnic origins were analyzed, resulting in the identification of only 5592 protein-coding genes with detectable mRNA transcripts common to all donors (Londin et al., 2014). From a large RNAseq study of platelets from 228 cancer patients and 55 normal donors, only 5003 (protein coding and non-coding) genes were detected. These fluctuations are likely due to various unknown factors, presumably including: age (Cowman et al. 2015), race (Edelstein, Simon, et al. 2013), RNA decay (Angénieux et al. 2016), gender (Cowman et al., 2015) and infection (Osman et al. 2015). Coupled with these fluctuations is the reported weak correlation between the transcriptome and proteome of platelets (Londin et al., 2014). Although over 85% of the proteome does not appear to vary very much between donors (Burkhart et al. 2012), <45% of detectable transcripts in platelets have an associated protein product (Londin et al., 2014).

It may provide insight to establish factors contributing to fluctuations in transcript counts and if such fluctuations impact PTES abundance. Three circRNAs from *CDR1* (Hansen et al., 2011; Memczak et al., 2013), *Sry* (Hansen et al., 2011) and *HIPK3* (Zheng et al. 2016) have been shown to be miRNA sponges, sequestering specific miRNAs away from their targets. However, miR-671 was shown to induce degradation of circCDR1 (Hansen et al., 2011), presumably through endonuclease activity, indicating that miRNAs may target PTES

transcripts. As miRNAs are abundant in platelets (Edelstein, Mckenzie, et al. 2013; Landry et al. 2009), there may be an observable impact upon circRNA abundance.

5.2 Aims

To further elucidate the global properties of PTES transcripts and infer their functional relevance, in this chapter, my aims were:

- To define the distribution of PTES in anucleate cells and compare their abundance to that of PTES in nucleated cells
- To identify platelet-specific PTES events
- To identify factors that may contribute to fluctuations in PTES biogenesis and abundance between platelet preparations

5.3 Results

To further elucidate the global properties of PTES transcripts and characterize their distribution in anucleate cells, I hypothesized that PTES transcripts are long lived and abundant in platelets, given their relative abundance in cytosolic fractions of nucleated cells. To that end, ribosome depleted RNAseq data from platelets preparations of 3 individuals (2 males, 1 female) and one platelets polyA+ sample from a single male donor were obtained from NCBI Gene Expression Omnibus database. These data were previously mined for non-coding RNA in Kissopoulou et al., (2013), thus, meet library size standards for identifying novel transcripts. In the absence of an equivalent dataset from red blood cells (RBC) to act as a comparator anucleate cell type, ribosome depleted RNA extracts from mature erythrocytes were sequenced, obtaining >100 million paired-end 100bp reads (see methods). These datasets were screened for PTES using PTESFinder v. 1, with default parameters. From the platelets total RNA samples, 33,829 distinct PTES transcripts were identified and are produced from 6198 genes (full lists in appendix 9.4). The highest number of transcripts was identified from the female platelets sample (Platelets_F), with 29,499 PTES transcripts, supported by 769,249 reads. On average, ~0.9% of reads screened support PTES events, contrasting <0.005% of reads observed per sample in the previous analysis (chapter 4), an ~240X increase in detectable PTES reads. A high overlap in identified PTES was observed, with ~74% of PTES identified from at least 2 platelets samples (Fig 5.1A). In the RBC sample, 12,380 transcripts were identified and supported by 82,516 reads, representing ~7 reads per PTES junction. Comparatively, ~85% of transcripts identified in RBC were also identified in at least one platelets sample (Fig 5.1B). Results from pairwise correlation analyses using PTES junction counts showed high correlation in PTES abundance between anucleate samples (Fig 5.1C), with the highest correlation coefficient (0.96) observed between Platelets_M2 sample and the female platelets sample (Platelets_F). Remarkably, this high degree of concordance is greater than observed for any pairwise comparison between nucleated cell types (chapter 4).

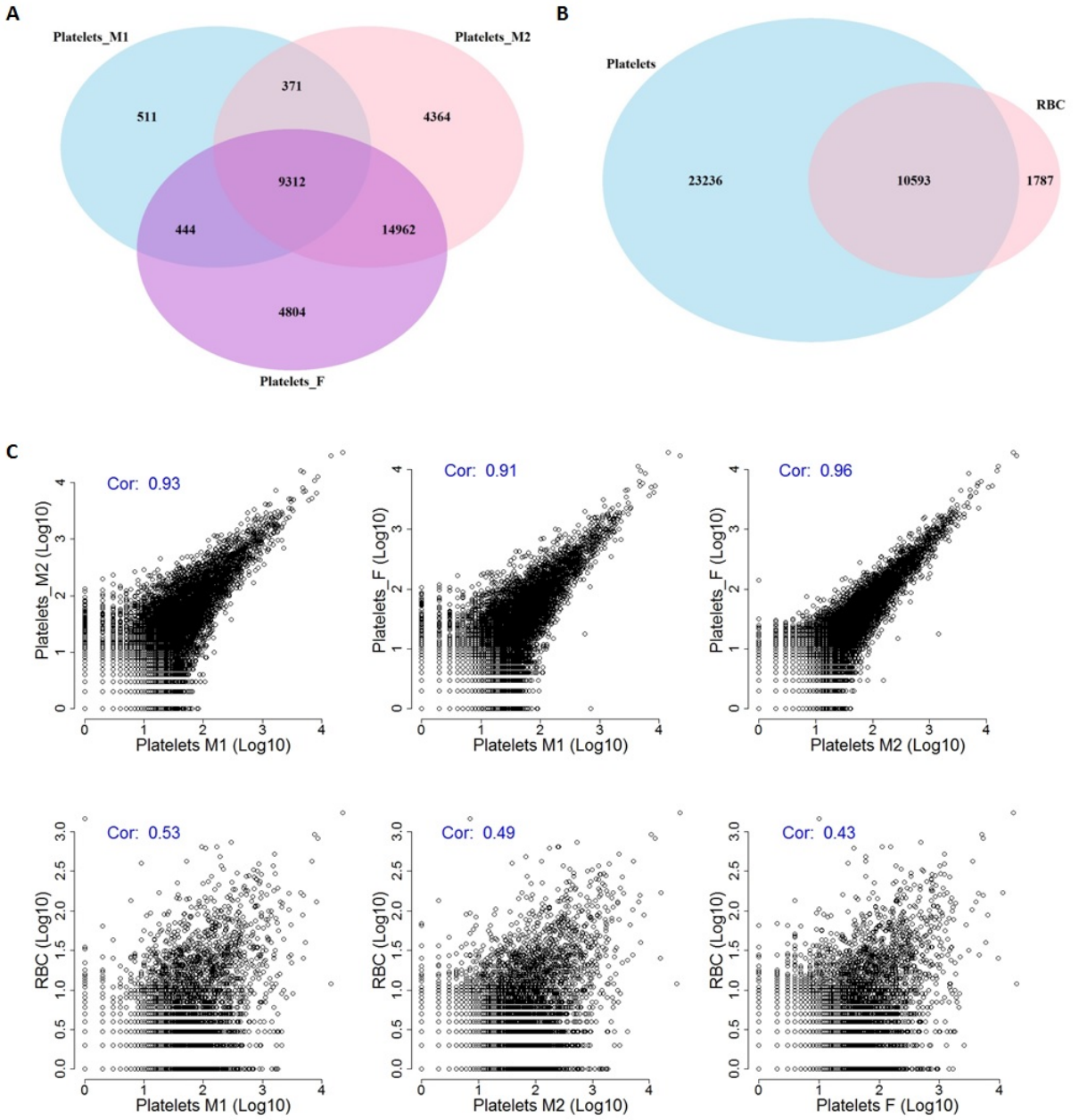


Figure 5.1. **Exploratory analysis of PTES in anucleate ribosome-depleted samples.** A -B) Venn diagrams showing overlap of PTES transcripts identified from 3 ribosome depleted platelets samples (Kissopoulou et al. 2013) and one sample from red blood cells. C) Pairwise correlational analyses of PTES from all four samples.

5.3.1 Most PTES transcripts identified in platelets are circular

From the single polyA+ platelets sample analyzed, 841 transcripts originating from 453 genes were identified, lower than observed from total RNA samples. Previous studies have inferred linear PTES transcripts from analysis of PolyA+ fractions (Al-Balool et al., 2011), as only linear transcripts are expected in such samples. However, many of the transcripts identified from the polyA+ sample are observed with higher read counts in the total RNA samples, suggesting enrichment in ribosome-depleted samples. Furthermore, some of these transcripts have been reported in previous studies and identified from RNase R treated samples.

Comparing identified transcripts to transcripts reported from RNase R digested samples (and within circbase.org), there is an overlap of 78.2% (658) with transcripts identified from the polyA+ sample, suggesting these transcripts to be circular.

Without ruling out the possibility of some of these PTES transcripts existing as both linear and circular molecules, a nucleotide composition analysis was performed to assess the likelihood that they are circRNAs pulled down during polyA+ selection. Comparing the number of adenosine residues for these transcripts to that of transcripts identified from total RNA samples only, I found that polyA+ PTES transcripts are significantly enriched for adenosine residues ($X^2 = 12317$, df = 1; p-value < 2.2×10^{-16}). For some transcripts, adenosine residues constitute ~60% of total nucleotide composition (examples in Fig 5.2). On the evidence of reported resistance to RNase R and their nucleotide composition, these transcripts are likely not linear and contaminate polyA+ samples. Furthermore, ~63% of transcripts identified from total RNA samples have been observed in previous studies (Fig 5.3) and from RNase R samples, suggesting their circularity.

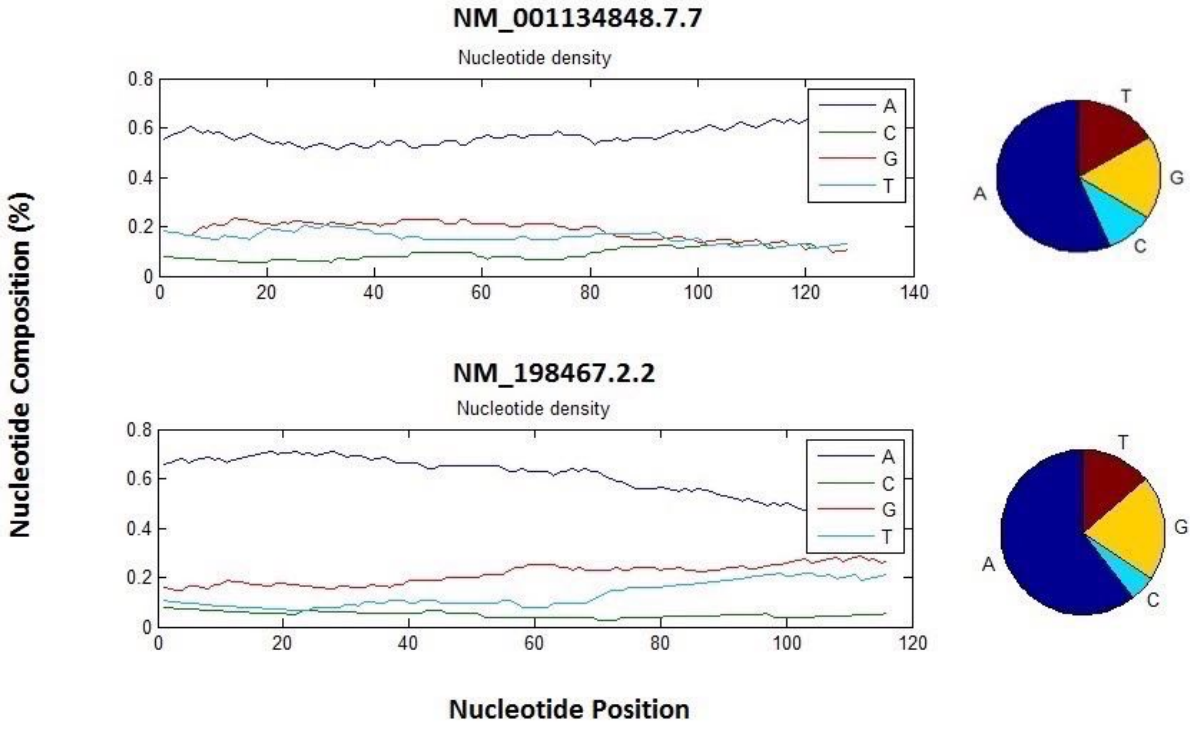


Figure 5.2. **Examples of A-rich circRNAs.** A) Nucleotide composition of two circRNAs, derived by sliding a 20bp window across respective circRNA sequences. Pie charts show relative proportion of each nucleotide in circRNA sequence. PTES ids are composed of RefSeq ID, donor exon order and acceptor exon order; circRNA sequences were generated by concatenating nucleotide sequence of each exon, consistent with PTES.

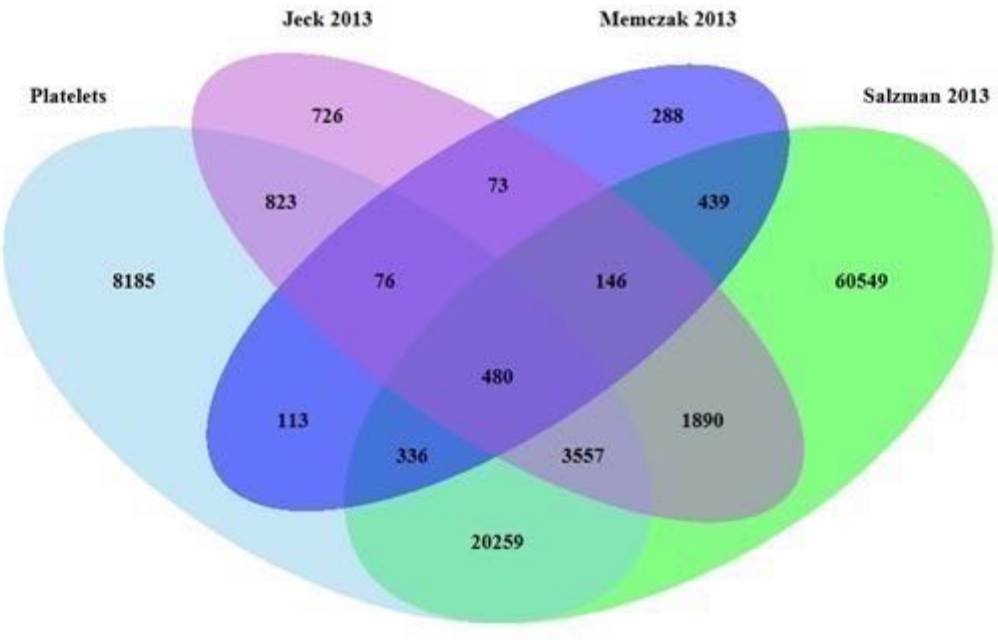


Figure 5.3. **Overlap of identified transcripts with published circRNA transcripts**. Distinct transcripts identified from all 3 platelets samples are compared with circRNA transcripts reported in Jeck et al., (2013), Memczak et al., (2013) and Salzman et al., (2013). Approximately 63% of PTES in platelets were reported in these studies. Figure taken from Alhasan et al., (2016)

5.3.2 CircRNAs are enriched in anucleate cells and expand the growing catalog of PTES transcripts

In addition to the high overlap of identified PTES transcripts with that of previous reports, additional unreported transcripts were identified from many known PTES producing genes. In platelets, over 600 genes have more than 20 different circRNA transcripts, underscoring PTES as a mechanism contributing to transcriptome diversity. One example is *XPO1*, a gene involved in RNA export pathways. Eleven circRNAs from this gene were previously identified from RNase R digested H9 ESC (Zhang et al., 2014); however, 38 additional transcripts were identified within the platelets samples analyzed (Fig 5.4).

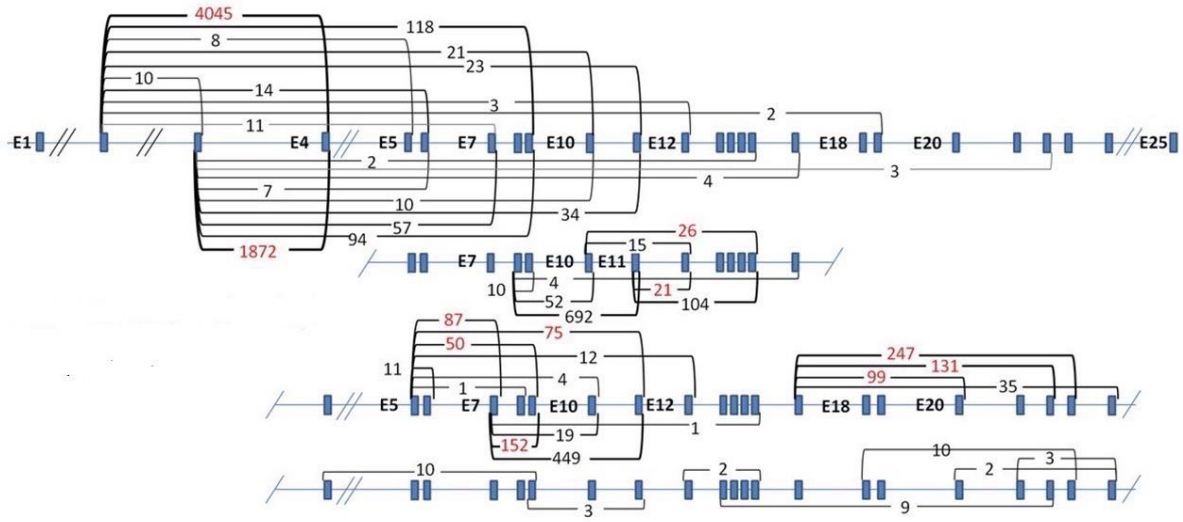


Figure 5.4. **Previously uncharacterized transcripts from *XPO1* locus.** Schematic diagram showing intron/exon structures of previously identified PTES from *XPO1* and PTES identified in platelets. Eleven (highlighted in red) were characterized in Zhang et al., (2014), additional 38 unreported transcripts were identified from platelets; read counts for each transcript is shown. Figure from Alhasan et al., (2016).

To extend comparisons to samples from nucleated cells, 12 ribosome depleted total RNA samples from various human cell lines, previously mined for ncRNAs were obtained (Nielsen et al., 2014). Additionally, 9 samples from 4 cell lines (H9 [Zhang et al., 2014], Fibroblasts [Jeck et al., 2013], K562 and GM12878 [Birney et al., 2007]) were also obtained and analyzed using PTESFinder.

Collectively, analyzed samples can be grouped into: anucleate cells, nucleated tissues, nucleated cell lines and RNase R digested nucleated cell lines. Strikingly, the highest number of PTES transcripts and PTES supporting reads were observed in anucleate cells (Fig 5.5). Undigested nucleated cell lines had the lowest number of distinct PTES transcripts and lowest mean circRNAs per gene (Fig 5.5). In nucleated tissues, the highest number of PTES transcripts were identified from brain (n = 4847). However, transcripts identified in platelets are supported by more read counts than nucleated tissues (p-value: 2.7×10^{-4}, Wilcoxon rank sum test). Furthermore, in each platelets sample, over 1000 transcripts are supported by over 100 reads, higher than in nucleated samples (Table 5.1).

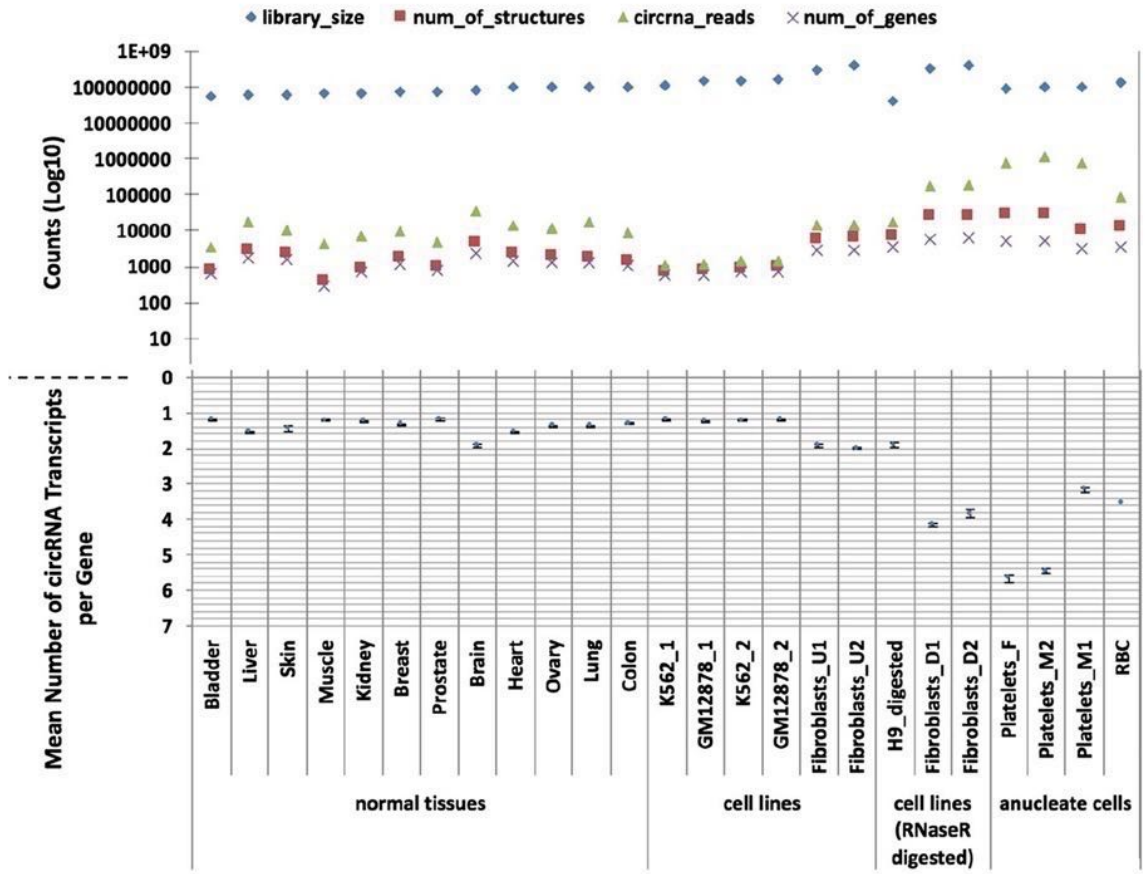


Figure 5.5. **Summary of PTES transcripts identified from human tissues and anucleate cells.** The number of identified circRNAs, supporting read counts, circRNA producing genes, mean number of circRNAs per gene and library sizes are shown for all samples analyzed. Samples are grouped in 4: normal nucleated human tissues, human cell lines, human cell lines digested with RNase R and anucleate cells. The number of circRNA supporting reads is significantly higher in anucleate samples than in others (p-value: 2.7 x 10^{-4}, Wilcoxon rank sum test). Figure and legend from Alhasan et al., (2016)

PTES Reads	Bladder	Brain	Breast	Colon	Platelets_M1	Platelets_F	Platelets_M2	H9_digested	Heart	Kidney	Liver	Lung	Muscle	Ovary	Prostate	RBC	Skin	Fibroblasts_U1	Fibroblasts_D1	Fibroblasts_U2	Fibroblasts_D2	GM12878_1	GM12878_2	K562_1	K562_2
>10	55	670	186	190	8459	7281	9605	227	235	196	308	521	136	263	71	1272	164	125	2797	141	3052	9	6	3	11
>100	0	19	3	5	1325	1197	1686	2	5	2	11	6	2	3	1	96	3	4	200	2	224	0	0	0	0
>1000	0	0	0	0	95	83	147	0	1	0	0	0	0	0	0	2	0	1	2	0	5	0	0	0	0

Table 5.1. **Frequency of reads per PTES junction.** Number of PTES transcripts identified from nucleated tissues and anucleate cells, with >10, >100 and >1000 PTES supporting reads

In an attempt to identify platelets-specific PTES transcripts, I generated a combined list of previously reported PTES transcripts in circbase.org, PTES identified in previous analyses and PTES from nucleated samples analyzed here. Comparing the genomic coordinates of PTES transcripts identified in >1 platelet sample, with my transcripts in the combined list, I identified 1260 novel transcripts that are seemingly platelet specific. Using read counts from the female platelets sample, novel PTES transcripts are supported by fewer reads than observed for

previously identified transcripts (p-value: 2.107 x 10^{-13}, Wilcoxon rank sum test), suggesting their rarity (or sampling bias) as a plausible reason for non-detection. Of note, however, are transcripts from *EFCAB13* and *BANK1*. Both have multiple previously unreported PTES transcripts (43 for *EFCAB13* and 23 for *BANK1* [Fig 5.6A-B]), that are supported by numerous reads and flanked by multiple *Alu* repeat elements that presumably aid their circularization. Expression estimates of linear transcripts from these genes are low in nucleated tissues, suggesting that they are platelets specific (Table 5.2), and that circRNAs from these loci result from increased transcriptional output. A screen for circRNAs from interleukin-1B, previously shown to undergo cytoplasmic splicing in platelets, did not identify any transcripts. Although platelets can become activated at room temperature (Maurer-Spurej et al. 2001), these samples were generated from resting platelets and may not contain cytoplasmic spliced products, as this is known to occur upon activation (Denis et al., 2005).

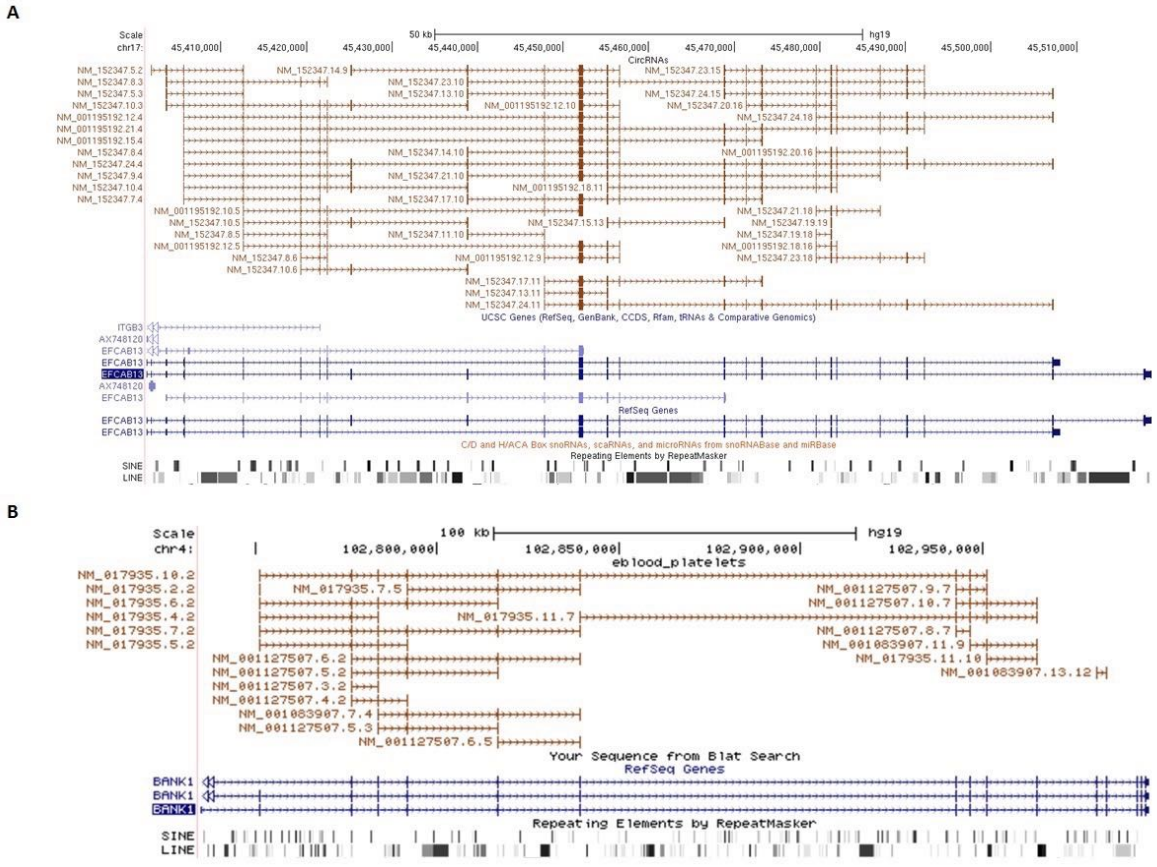


Figure 5.6. **Multiple PTES transcripts from *EFCAB13* and *BANK1***. Two examples of genes with multiple PTES transcripts and highly expressed in Platelets relative to other tissues. Multiple PTES transcripts from the same locus underscore the diversity of eukaryotic transcripts and are easily detectable in the platelets samples analyzed in this study.

Gene	Bladder	Brain	Breast	Colon	Fibroblasts_D1	Fibroblasts_U1	Fibroblasts_U2	Fribroblasts_D2	GM12878_1	GM12878_2	H9_digested	Heart	K562_1	K562_2	Kidney	Liver	Lung	Muscle	Ovary	Platelets_F	Platelets_M1	Platelets_M2	Prostate	RBC	Skin
BANK1	69	25	159	230	3	27	4	19	2	3	1	1	2	0	0	94	791	0	16	21575	15772	27079	127	0	68
EFCAB13	226	89	461	523	152	159	46	31	17	12	6	195	4	17	212	474	756	0	538	105010	86404	162016	148	669	306

Table 5.2. **Raw counts of platelets specific genes**. Raw counts of EFCAB13 and BANK1 from all samples, showing the highest number of counts observed in platelets samples

5.3.3 Reads from circRNA producing exons are enriched in Platelets

Previous studies have relied on flanking canonical junction counts to normalize PTES junction abundance for quantitation (Salzman et al., 2012). For instance, raw counts for a hypothetical PTES between exons 6 and 3, are often normalized using raw counts of canonical junctions 5-6 and 2-3. However, as shown for *XPO1*, EFCAB13 and *BANK1*, PTES transcripts can overlap and contribute counts to canonical junction counts used for normalization. To mitigate against this potential confounding factor, all PTES transcripts identified from all four groups of samples were first pooled and used to determine exons internal and external to circRNAs. This information was then used in computing expression estimates of circRNA exons ($RPKM_I$ - representing aa measures of linear and circular transcripts) and external exons ($RPKM_E$), representing a measure of linear transcripts only (see 2.6.4 for details). Results show that reads from circRNA exons are on average ~12X more abundant in platelets than in nucleated tissues and ~5X more abundant than RNase R digested samples (Fig. 5.7A).

To estimate the overall contribution of circRNAs to total transcriptional output of each gene, ratios of $RPKM_I$ / ($RPKM_I$ /$RPKM_E$) were derived (appendix 9.4). An enrichment analysis using these ratios was performed to identify PTES producing genes enriched for circRNAs in platelets. Using the Wilcoxon rank test, ratios derived for all three platelets samples were compared to ratios for 12 nucleated tissues. After correcting for multiple testing using the Benjamini-Hochberg protocol at FDR of 0.01, 3162 of 8041 genes tested reached significance, indicating enrichment in platelets relative to nucleated tissues. Furthermore, using genes with RPKM > 1 in platelets showed that for most genes enriched for circRNAs, the contribution of circRNA exons to total transcription was >80% in platelets, higher than observed for nucleated tissues (>60%). For 457 genes, the contribution of circRNA exons to overall transcriptional output exceeded 99% in platelets. I then estimated the magnitude of circRNA enrichment in platelets, relative to nucleated tissues (Fig. 5.7B-C; appendix 9.4). When mean RPKM ratios derived for both platelets and nucleated tissues were compared, I observed an average enrichment in circRNA exons of 12.7X for all genes and 22X for 3162 identified as significantly enriched. *TMEM181*, for instance, has 17 exons, 12 of which are internal to one or more of the 12 PTES transcripts identified from this gene. In platelets, the vast majority of reads observed

for this gene originate from the 12 exons internal to circRNAs, resulting in 3590X enrichment (Fig 5.7C).

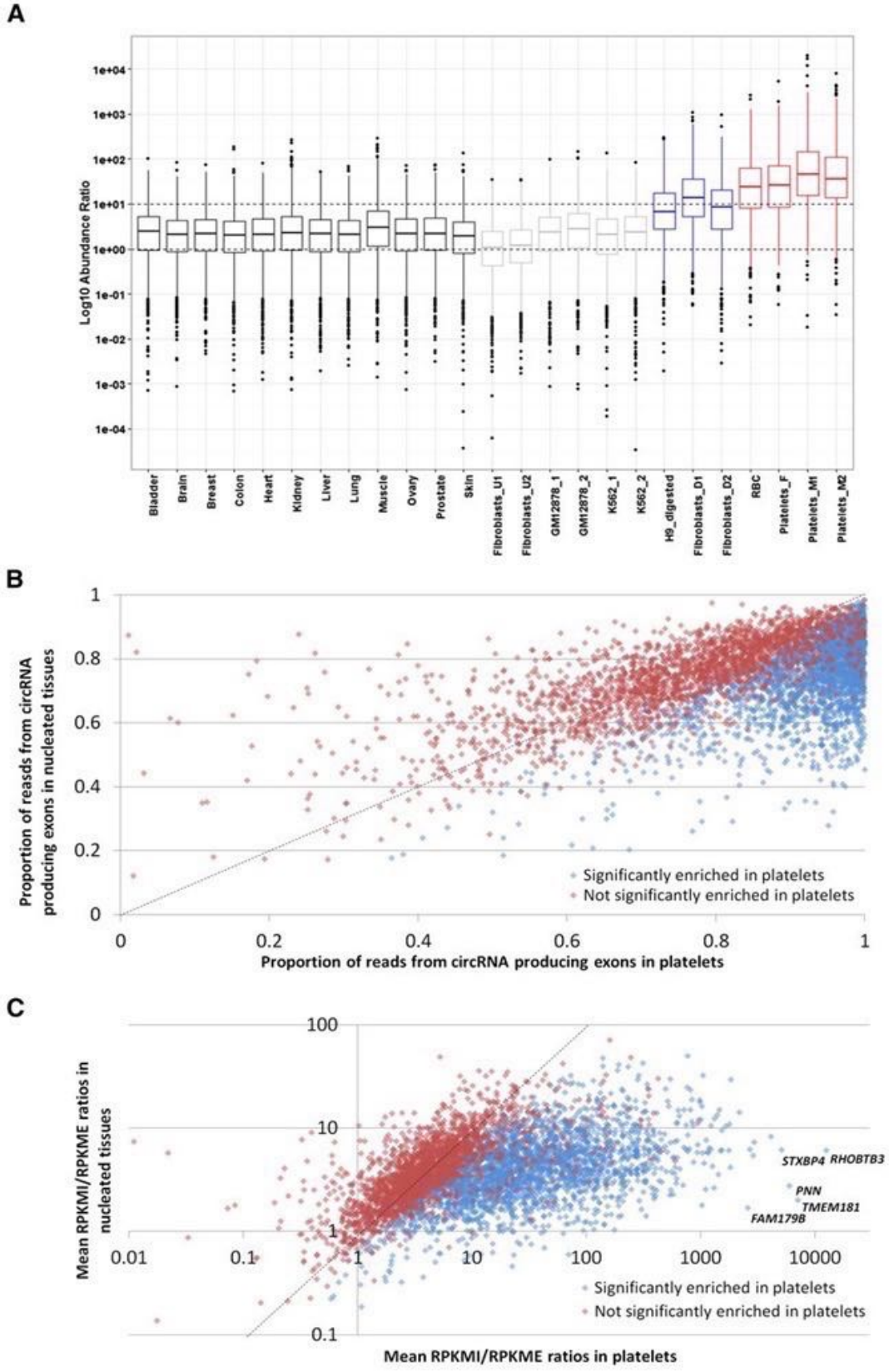


Figure 5.7. **Differential read depth defines genes with significant circRNA enrichment in platelets.** (A) Box and whisker plots showing the ratio of RPKM from circRNA producing exons (RPKMI) to RPKM from exons that do not produce circRNAs (RPKME) for all genes in each sample. The median and upper and lower quartiles are shown, with outliers as solid

circles. (B) The proportion of reads from circRNA producing exons averaged across all nucleated samples (y-axis) and platelets (x-axis). (C) Fold enrichment of reads from circRNA producing exons in platelets relative to nucleated tissues. All genes with an average RPKMI >1 in platelets and expressed in 8 or more nucleated tissues are shown. Blue, genes significantly enriched in platelets; red, genes not significantly enriched in platelets. The data points corresponding to the 5 most enriched genes are indicated. The slope x = y is shown as a dashed line. Figure and legend from Alhasan et al., (2016).

The observed abundance of circRNAs in platelets was confirmed using qPCR. Eleven previously confirmed circRNAs (including 2 each from *MAN1A2* and *PHC3*) (Al-Balool et al., 2011) were assayed by Dr. Alhasan (Newcastle University, UK), and their expression compared to that of associated canonical junctions, sharing the same probes (Fig 5.8A). The qPCR results in Fig 5.8B-C shows that these PTES transcripts are highly abundant in platelets and RBC, relative to nucleated tissues, registering between 4 & 10 cycles before their linear counterparts for 9 of the circRNAs assayed. This suggests an enrichment of ~16X to 1000X relative to their linear counterparts, significantly higher than observed in nucleated tissues (p-value: 2.7×10^{-12}, Wilcoxon rank sum test). The circularity of 7 PTES transcripts was experimentally confirmed by qPCR (Fig. 5.8D), after treating samples with RNase R to degrade linear molecules. Increase in expression estimates is noticeable for all assays, with the largest changes observed in nucleated samples, indicating that these samples contained a greater proportion of linear molecules than the platelets samples.

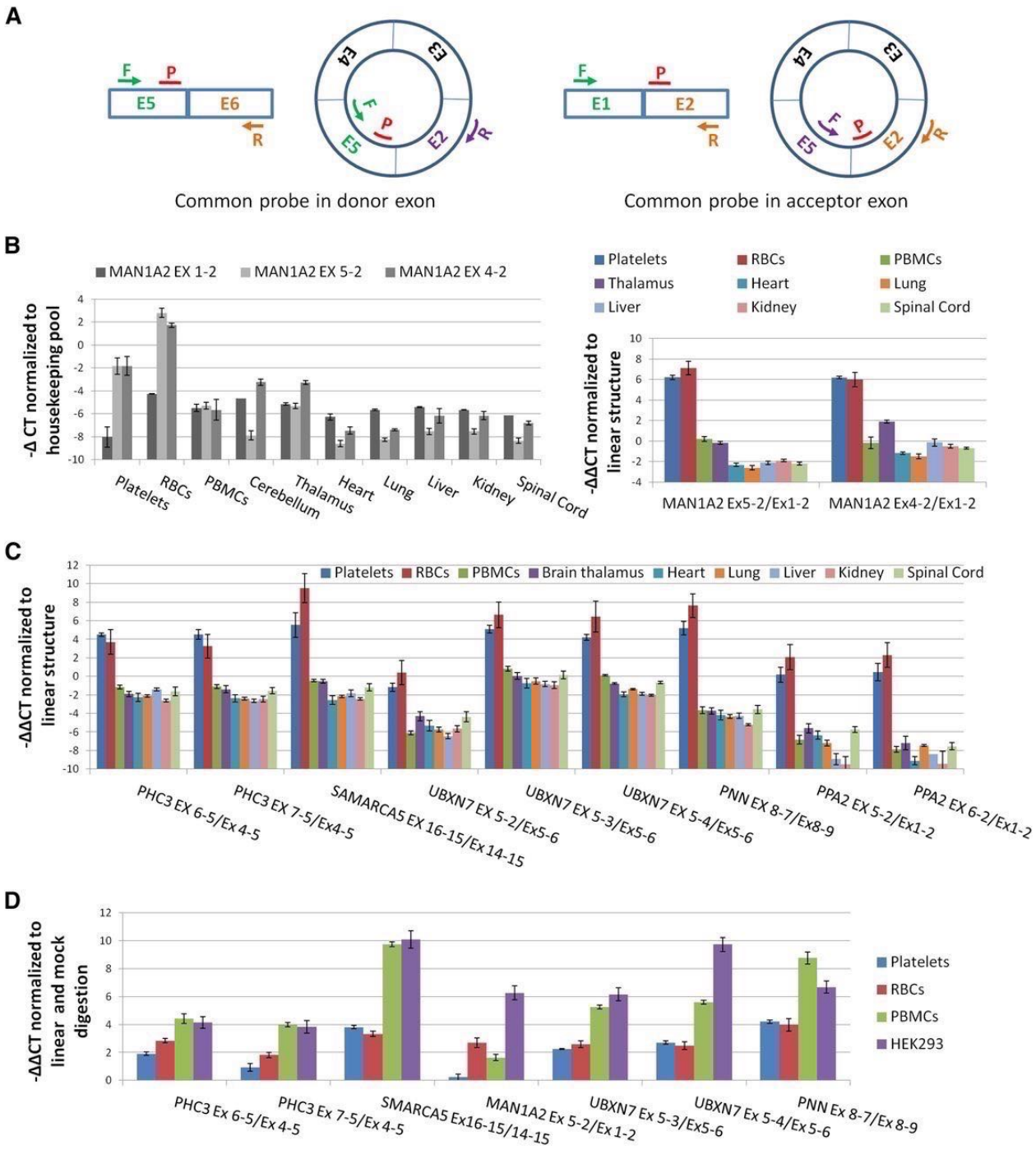


Figure 5.8. **Confirmations of circRNA abundance and resistance to RNase R.** A) Schema of qPCR assays using E5-E2 circRNA as an example. All assays use a common reported probe and use either an exon downstream of the circRNA to assay linear expression (probe in donor exon) or an exon upstream (probe in acceptor exon). (B) (Left) Expression levels (−ΔCT values) of linear (Ex1-2) and circular (Ex5-2 and Ex4-2) *MAN1A2* transcripts relative to housekeeping pool. (Right) Expression levels of circRNAs relative to linear RNAs from the same loci normalized to housekeeping pool (−ΔΔCT values). (C) Expression of 9 circRNAs relative to linear forms from the same loci, normalized to housekeeping pool (−ΔΔCT values). (D) Change in CT values of circRNAs relative to linear forms from the same loci in RNase R digested RNAs, normalized to mock digested RNAs (−ΔΔCT values). Templates, circRNAs, and linear forms assayed are indicated. Figure and legend from Alhasan et al., (2016).

5.3.4 RNA degradation may be reason for PTES abundance in anucleate cells

Why are PTES transcripts enriched in anucleate cells? In the absence of steady transcription and export from the nucleus, expression levels of mature mRNAs decrease (Angenieux et al., 2016). Various RNA decay pathways include exonucleases that act to de-cap, de-adenylate and degrade mature mRNAs (Houseley & Tollervey 2009; Schoenberg & Maquat 2012). Using *in silico* and *in vitro* methods, I assessed RNA decay in anucleate cells. First, for each sample, I computed the percentile base coverage of every RefSeq annotated transcript observed in that sample. For instance, a transcript with internal spliced size of 8000bp will generate 100 segments of 8bp in sizes. The number of nucleotides within each segment covered by at least one read is used to compute the coverage for that segment. For all transcripts the mean percentile coverage was derived and compared across samples (Fig 5.9A). Results show that anucleate cells have the least mean percentile coverage across transcripts, suggesting RNA decay or possible sampling bias. Unlike nucleated samples, there is a noticeable reduction in nucleotide coverage towards the 3' termini of transcripts in each anucleate sample. This is consistent with exonuclease activity and likely contributes to the magnitude of circRNA enrichment observed using expression estimates of exons external to circRNAs, as terminal exons typically are not involved in PTES. Second, as platelets originate from megakaryocytes, we reasoned that expression differences between both cell types will be decay rate dependent. To assess this, I identified and obtained a polyA+ sample from megakaryocytes. PolyA+ samples from both cell types were then analyzed and expression estimates of each gene compared across samples. When the ratios of expression estimates obtained from both samples are compared to published half-lives of respective genes (Friedel et al. 2009), a significant correlation ($r = 0.17$, p-value: 3.35×10^{-44}; Fig 5.9B) between both measures is observed. This suggests that the expression of many transcripts decreases in the platelets relative to their progenitor cells, in a decay rate dependent manner.

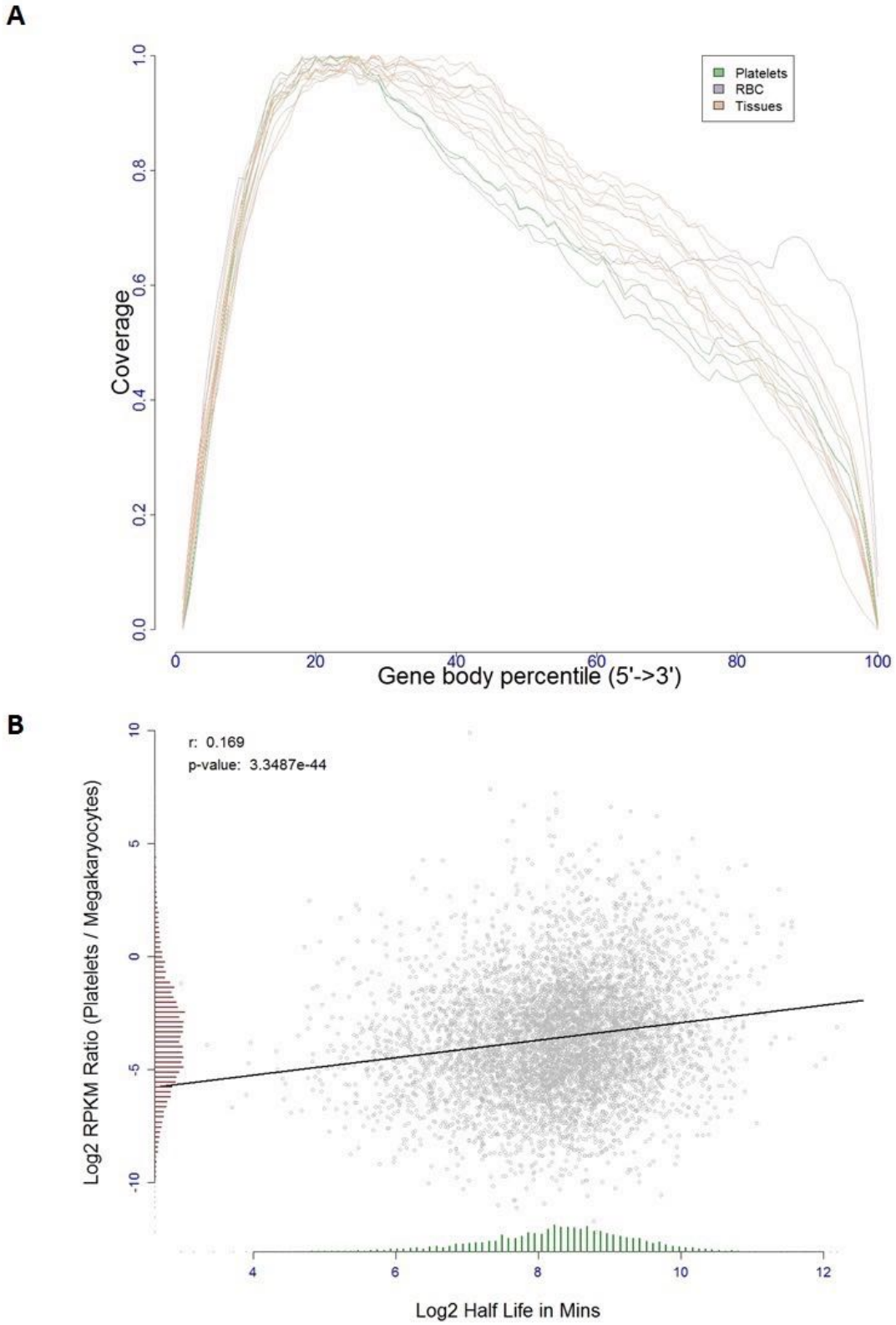


Figure 5.9. ***In silico*** **decay analysis in platelets.** A) Mean read coverage across transcripts identified in Platelets, RBC and nucleated tissues. Proportion of bases covered by reads in sequence segments (n = 100) of each transcript is averaged and shown for all samples. B) Correlation of the change in gene expression between megakaryocte polyA+ RNA and platelet polyA+ RNA (y-axis), and mRNA half-life estimates (x-axis). Distributions of both are shown as histograms on each axis. Figure 5.9B from Alhasan et al., (2016)

To confirm these results experimentally, plasma rich platelets from 3 individuals were incubated at 37°C for 72 and 96 hours by Dr. Alhasan (Newcastle University, UK), monitoring the abundance of 4 housekeeping genes and 7 circRNAs, along with their cognate linear transcripts (Fig. 5.10). Results show a reduction in abundance of housekeeping genes, registering 4 to 6 cycles later in decay time series, relative to the 0-hour time point. There is an equivalent reduction in linear transcripts assayed. However, circRNAs are enriched 2 to 6-fold relative to linear RNAs after incubation (p-value: 2×10^{-3} for comparisons of 0 vs 72hrs and 0 vs 96hrs; Wilcoxon rank sum test).

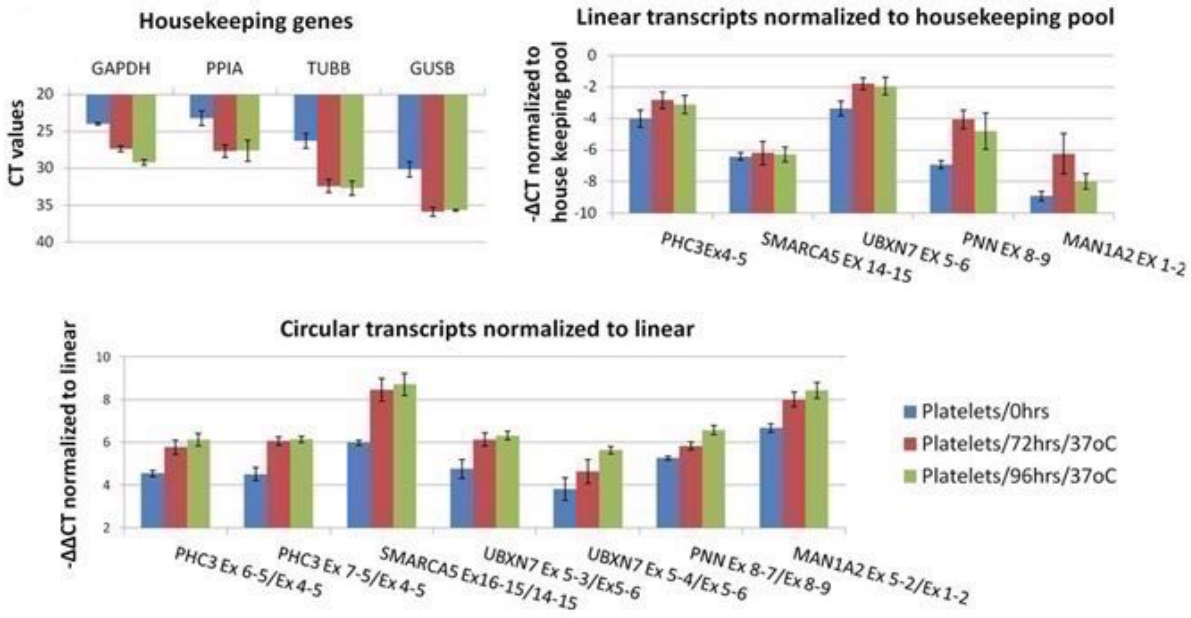


Figure 5.10. **Degradation of platelet RNA.** qPCR analysis of differential decay of linear and circRNAs in platelets following incubation at 37°C for 0, 72, and 96 hours. Data from 3 biological replicates are shown. (Top left) Expression levels of housekeeping genes (CT values). (Top right) Expression levels of linear structures from 5 circRNA-producing genes relative to the housekeeping pool (−ΔCT values). (Bottom) Expression levels of 7 circRNAs relative to linear transcripts from the same loci, both normalized to the housekeeping pool (−ΔΔCT values). Figure and legend from Alhasan et al., (2016)

5.3.5 RNA secondary structure, GC content and miRNA binding sites may contribute to circRNA stability

The stark difference in number of PTES transcripts and PTES supporting reads observed in the male platelets is noteworthy. In one platelets sample (Platelets_M1), only 10650 transcripts were identified, less than half the number observed in Platelets_M2, despite comparable library sizes. Abundance ratios computed using RPKM ratios, show the sample with fewer PTES transcripts to be significantly enriched for reads within circRNAs and deplete of reads in exons external to circRNAs (Fig 5.7A above), suggesting RNA decay of linear transcripts. However, many factors including read quality can contribute to the difference in PTES transcripts

identified between both male platelets samples. Per base sequence quality analysis of Platelets_M1, showed sub-optimal sequence quality in the last ~15bp of most reads, dissimilar from the other 2 platelets samples (Fig 5.11). As PTESFinder relies on generating short anchor reads from termini of reads, this observation is likely to cause the non-detection of some PTES transcripts. To assess this, I trimmed 20bp from all reads in this library and reanalyzed using PTESFinder. This correction only resulted in the identification of 33 additional PTES transcripts and fewer PTES supporting reads. When the impact of filters within PTESFinder is compared for all three samples, Platelets_M1 is the least affected sample, with only 38.07% of reads excluded by filters (Table 5.3), suggesting that the observed difference in identified transcripts in not due to poor sequence quality.

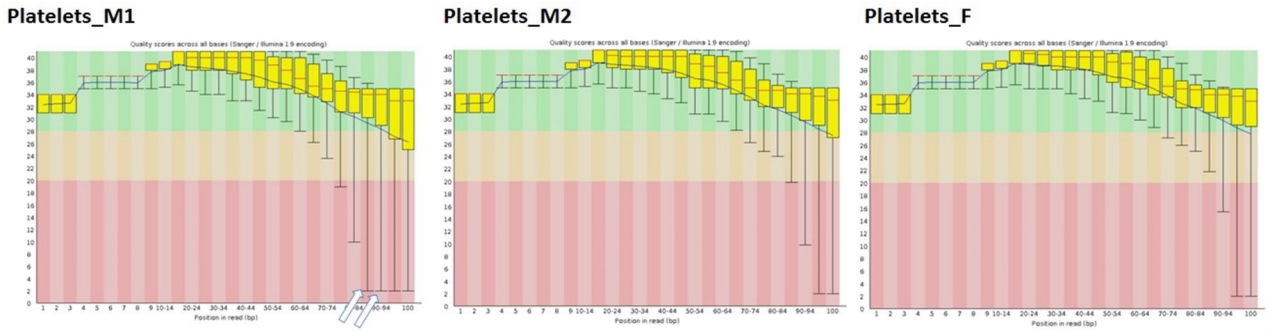


Figure 5.11. **Sequence quality analysis of Platelets samples.** Merged FASTQ reads for each Platelet sample were analyzed using FASTQC. Aggregate qualities of base calls across all reads are shown for each sample. Highlighted (white arrows) are nucleotide positions with sub-optimal base calls, differing from other samples.

Samples	PTES identified	PTES reads	Reads before filter	Filtered Reads (%)
Platelest_M1	10650	790098	1275892	38.07
Platelets_M1TR	10683	767910	1221396	37.13
Platelets_M2	29035	1152301	1861087	38.08
Platelets_F	29522	769249	1260318	38.96

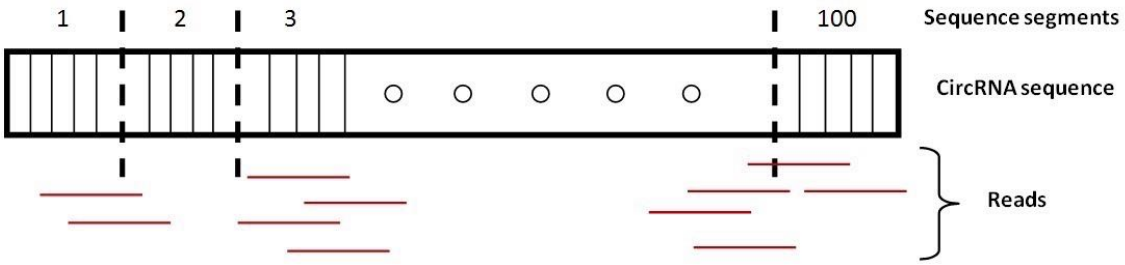
Table 5.3. **Re-analysis of Platelets samples.** Platelets_M1 sample was re-analyzed using PTESFinder after trimming 20bp off right ends of reads, to remove nucleotides with sub-optimal base call quality. This correction resulted in the identification of additional 33 PTES transcripts.

Sampling bias and possible technical errors may contribute to this effect, where some PTES transcripts are inadvertently affected. Indeed, when Platelets_M2 was subsampled without replacement using SEQTK (https://github.com/lh3/seqtk), generating 10 samples each containing 25% or 75% of library size, and screening for PTES; the number of identified PTES transcripts fluctuates. Interestingly, at either level of subsampling, a similar number of PTES transcripts was consistently identified across samples. It is also notable that even at 25% subsampling level, the numbers of identified transcripts are approximately 80% higher than observed in Platelets_M1 (Table 5.4), suggesting that reduced number of identified transcripts

is not solely due to sampling. Sampling cannot explain the observed enrichment of circRNA exons relative to non-circRNA exons (Fig 5.7A). It is however conceivable that some circRNAs in Platelets_M1 were undergoing degradation, negating detection. To investigate the intactness of circRNAs in both male samples, I pooled transcripts identified from both samples and generated their full spliced sequences (constructs) *in silico*. For instance, sequence for a circRNA with junction between exons 4 and 2, comprised of full sequence of exons 2, 3 and 4 concatenated at splice junctions. Full length reads were then mapped to these constructs; computing percentile read coverage as depicted in Figure 5.12.

Sample	Percent of Library Size	Identified Structures
Platelets_M2_1	25	18916
Platelets_M2_2		18921
Platelets_M2_3		18926
Platelets_M2_4		19004
Platelets_M2_5		18940
Platelets_M2_6		18834
Platelets_M2_7		18959
Platelets_M2_8		18908
Platelets_M2_9		18886
Platelets_M2_10		18978
Platelets_M2_1	75	27222
Platelets_M2_2		27168
Platelets_M2_3		27230
Platelets_M2_4		27230
Platelets_M2_5		27174
Platelets_M2_6		27217
Platelets_M2_7		27247
Platelets_M2_8		27209
Platelets_M2_9		27204
Platelets_M2_10		27236

Table 5.4. **Identification of PTES transcripts from sub-sampled Platelets_M2 sample.** Platelet_M2 sample were sub-sampled at 25% and 75% of total library size and screened for PTES transcripts. Sub-sampling affected number of identified transcripts by ~35% and ~7% respectively.



0.8 0.4 1.0 1.0 Percentile Coverage

Comparisons

Sample 1	0.8	0.4	1.0		1.0
Sample 2	0.7	0.5	0.6		0.2
Difference	0.1	-0.1	0.4		0.8

Wilcoxon rank sum test

Figure 5.12. **Methodology of percentile coverage comparisons.** Reads mapping to nucleotide sequence of each circRNA were used to derive percentile coverage. For each of the 100 segments of the full sequence, the number of nucleotides covered by reads is divided by the size of the segment. For instance, the first sequence segment depicted here has 4 of its 5 nucleotides covered by reads, resulting in coverage of 0.8. Sample-level comparison of percentile coverages for each circRNA was performed using Wilcoxon rank sum test. The difference in percentile coverage between samples is derived for all circRNAs.

Comparing percentile coverage for each transcript across both male samples, it is striking that for virtually all transcripts, there is a noticeable difference in percentile coverage between both samples (Fig 5.13A). The same pattern is observed when Platelets_M1 is compared to Platelets_F (Fig 5.13B), but differs from comparisons between Platelets_M2 and Platelets_F (Fig 5.13C). The distribution of percentile coverage differences shows that, for most transcripts, more nucleotides are covered by reads and are detectable in Platelets_M2 than in Platelets_M1 (Fig 5.13D). It is also striking that, comparisons between Platelets_M1 and subsamples of Platelets_M2 (Fig. 5.13E-F), show similar patterns, further indicating that the percentile coverage metric is not significantly affected by sampling.

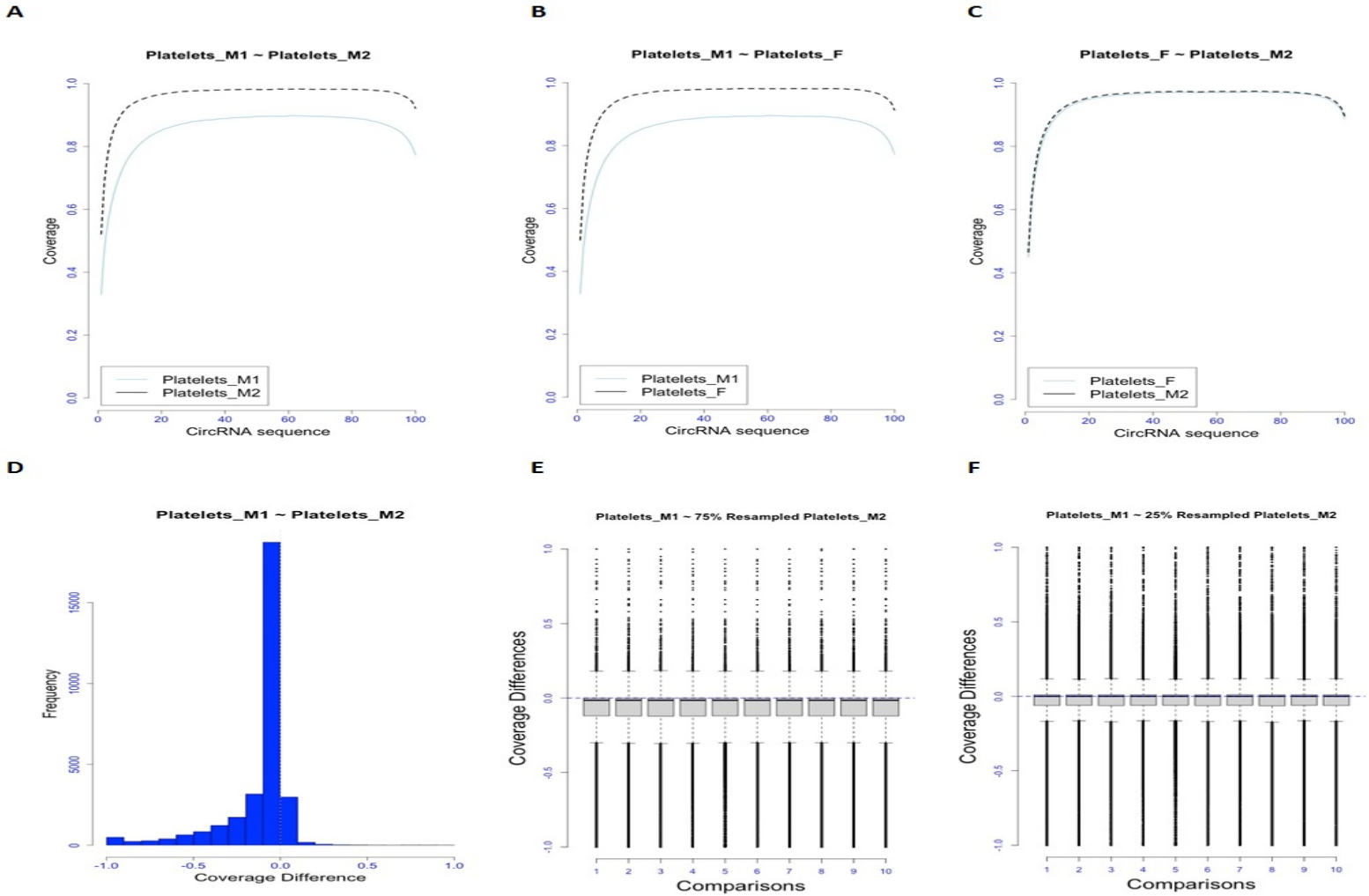


Figure 5.13. **Comparisons of percentile coverage across circRNA sequences.** Mean percentile coverage across circRNA transcripts identified from A) Platelets_M1 & Platelets_M2 B) Platelets_M1 & Platelets_F and C) Platelets_F and Platelets_M2. D) Differences in mean percentile coverage between both male platelets are shown E - F) Differences in mean percentile coverage between Platelets_M1 and 25% or 75% subsampled Platelets_M2

To identify transcripts with significant differences in coverage, I performed Wilcoxon rank tests using derived percentile coverage for each transcript. From ~31000 transcripts tested (appendix 9.4), 11,400 transcripts significantly differed in percentile coverage and have more nucleotides covered by reads in Platelets_M2, after correcting for multiple testing using BH method. Only 67 transcripts were found to have more nucleotides covered by reads in Platelets_M1. These results suggest that most PTES transcripts are more intact in the second platelets sample, in addition to the higher number of transcripts identified from that sample. Figure 5.14 shows examples of transcripts, highlighting differences in read coverage across transcripts between both male platelets samples. In these examples, the circRNAs shown are intact in Platelets_M2, but there are gaps in read coverage in Platelets_M1.

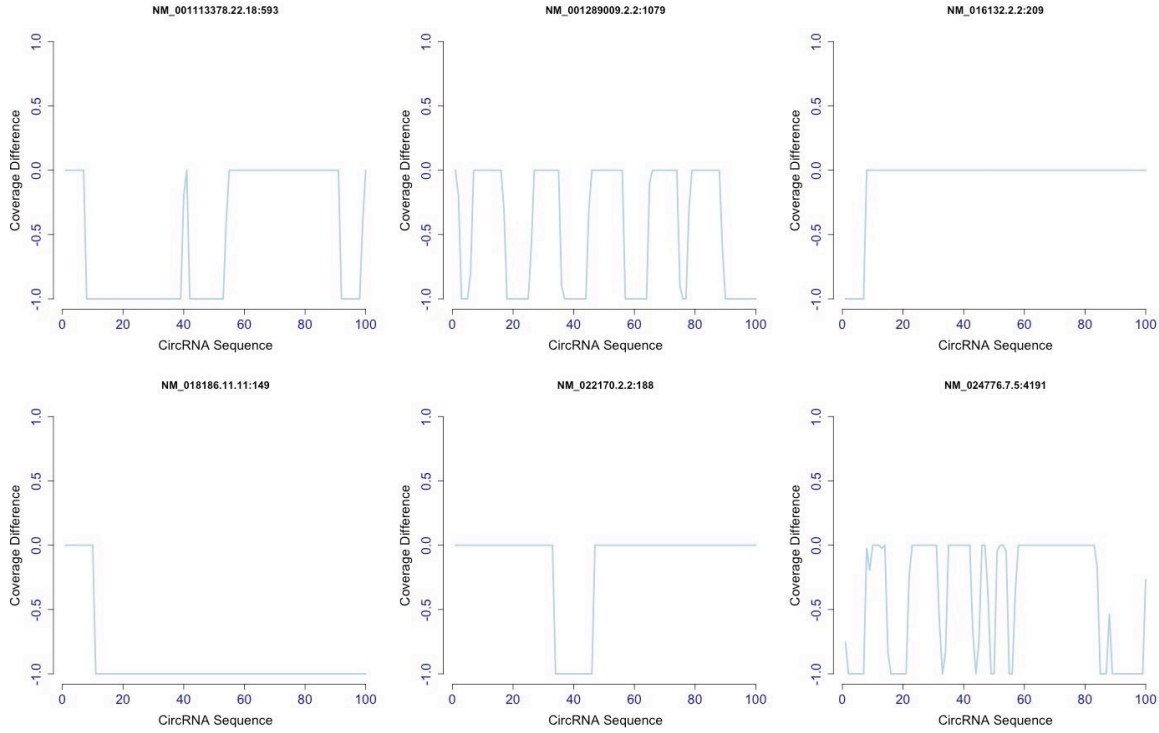
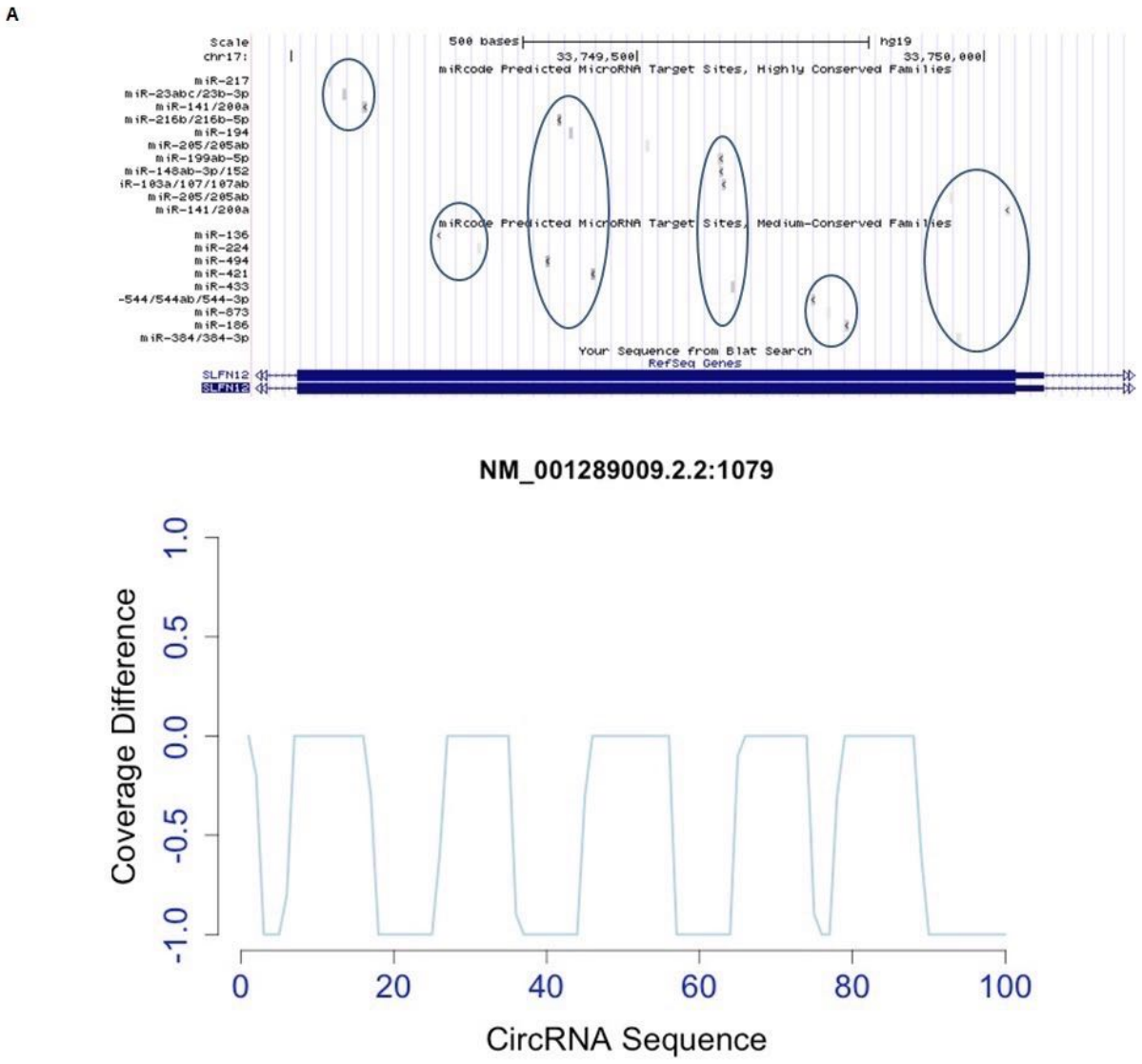


Figure 5.14. **Examples of transcripts with non-uniform read distributions in Platelets_M1 sample**. Difference in percentile read coverage between two male samples is shown for 6 transcripts. Transcript ids are composites of RefSeq id, donor exon, acceptor exon and size of circRNA exon. Each segment in circRNA sequence has a size corresponding to the transcript size divided by number of segments (100 in all cases). For all sequence segments, differences were derived by subtracting the coverage in Platelets_M2 from that of Platelets_M1, thus, differences favour higher coverage in Platelets_M2, for all examples shown.

In an attempt to identify features that may affect circRNA stability and potentially contribute to these observations, I compared the coordinates of pooled PTES transcripts to published miRNA binding sites (Jeggari et al., 2012), reasoning that miRNAs may act to suppress circRNAs. RNA degradation induced by miRNAs depends on full complementarity between recognition site on target RNA and the miRNA seed sequence. In some cases, however, the target RNA is not degraded, due to partial complementarity, resulting in translational

suppression for mRNA targets. Correlational analysis of the number of miRNA binding sites and difference in coverage (intactness) of circRNAs in both male platelets resulted in a weak but significant correlation (r: 0.06, p-value < 2.2 x 10^{-16}), suggesting that intactness and subsequent stability may be affected by miRNA activity. Notably, for some transcripts, the highest difference in coverage occur in regions with miRNA binding sites, highlighting the possibly role of miRNAs on circRNAs (Fig 5.15).

I then examined the effect of respective nucleotide composition on intactness. Results show that differences in coverage negatively correlate with GC content (r: -0.15, p-value < 2.2 x 10^{-16}). This is consistent with previous reports of the effect of GC on RNA stability (Wan et al., 2012). In a nutshell, some PTES transcripts not identified from Platelets_M1 or with variation in percentile coverage between samples, are characterized by numerous miRNA binding sites and lower GC content.



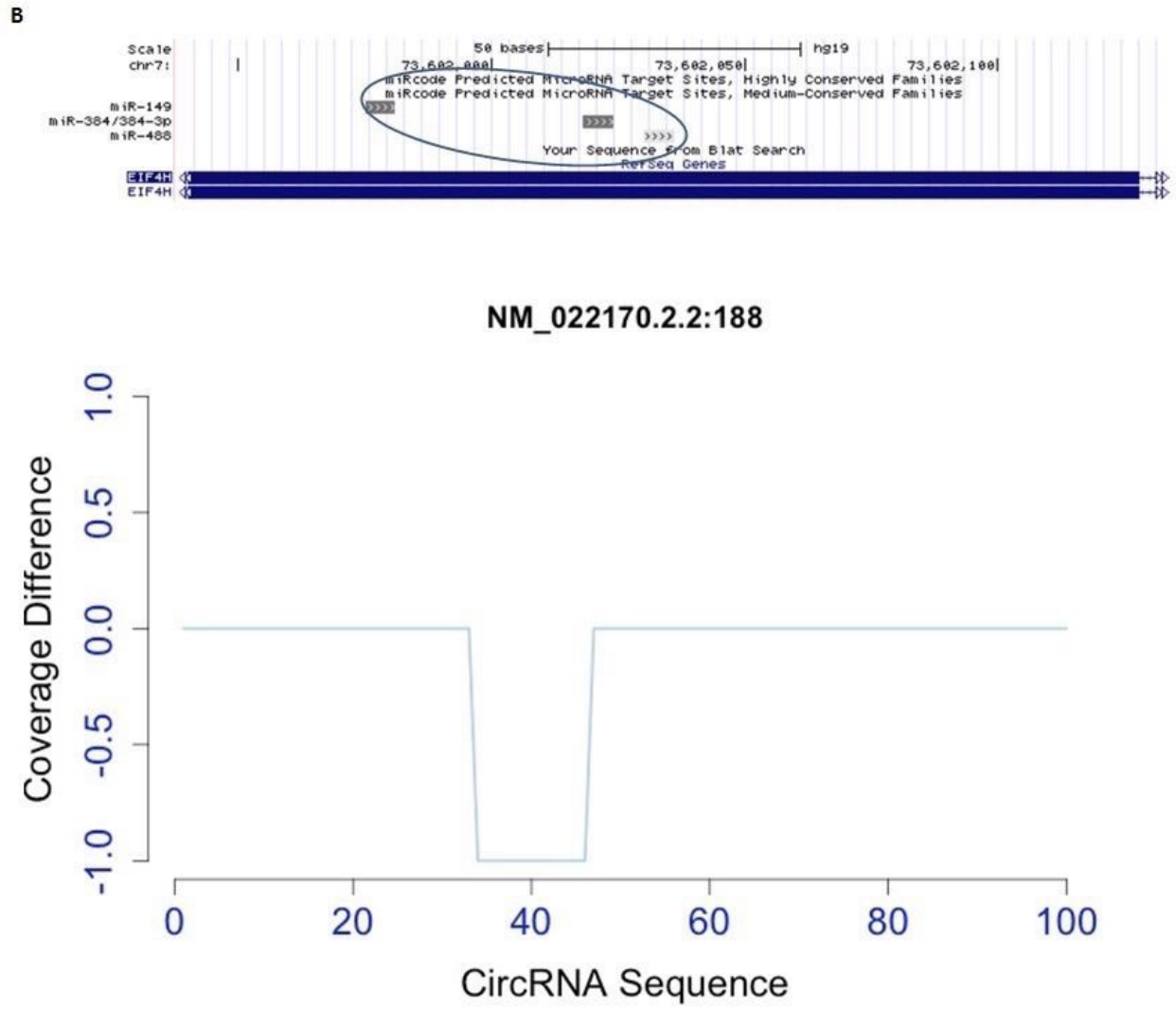


Figure 5.15. **Overlay of MiRNA binding sites on mono-exonic circRNAs.** A -B) Examples of single exon circRNAs with predicted miRNA binding sites overlaid on sequence segments. The size of the circRNA is appended to each circRNA ID, 1079bp and 188bp respectively.

5.4 Discussion

Transcripts with rearranged exon order relative to the genomic context have now been identified in various nucleated cell lines and tissues, but their identification and distribution have not been explored in anucleate cells. In the last chapter, ~9% of identified PTES transcripts were shown to be significantly enriched in the cytosol of nucleated cell lines, relative to the nuclei. Various reports have also identified circRNAs from cell-free RNA extracts, further suggesting that their resistance to exonuclease activity confers stability and their enrichment in cellular space. This singular premise highlights platelets as an ideal 'closed system', in the absence of steady transcription, for characterizing PTES transcripts. Platelets are highly specialized cells, devoid of genomic DNA and inherit their transcriptome from their megakaryocyte progenitors. In addition to being a snapshot of the megakaryocytic cytosol prior to release, platelets can accept exogenous RNAs from other cells, undergo cytoplasmic splicing and are enriched for incompletely processed miRNAs, underscoring the complexity of their transcriptomes.

To further characterize the distribution of PTES transcripts in cellular space and time, and identify any platelet-specific PTES transcripts, I analyzed 3 ribosome depleted total RNA samples from 3 donors and 1 polyA+ sample. From total RNA samples, ~34000 distinct PTES transcripts were identified, the highest reported from any cell line or tissue. Comparisons with 12 human tissues and 4 cell lines showed that the highest number of PTES transcripts, supporting reads and PTES producing genes were identified from platelets. Results from my analysis identified ~8000 previously unidentified PTES events, expanding the growing repertoire of characterized PTES transcripts. Within platelets, over 600 genes produce more than 20 distinct PTES transcripts, many of which were previously unreported. As an example, 11 circRNAs were previously identified from *XPO1*, in platelets; an additional 38 circRNAs were identified, highlighting the high resolution of datasets analyzed. For each locus, comparing the abundance of reads from circRNA exons to exons external to any identified circRNA, all three platelets samples were found to be ~12X and ~5X more enriched for reads within circRNA exons than nucleated samples and RNase R digested samples respectively. Further enrichment analysis identified ~3000 genes to be significantly enriched for reads within circRNA exons, in some cases, >90% of reads emanate from circRNA exons. A similar observation was made in mature erythrocytes, indicating that circRNA enrichment may be a common feature in anucleate cells. *In vitro* analysis of 11 circRNAs confirmed circRNA enrichment in anucleate cells, relative to nucleated cells, including megakaryocytes.

The observed abundance of reads from circRNA exons in platelets is likely due to differential decay of linear RNA molecules and circRNA stability. As reads from exons external

to circRNAs originate from linear molecules, their depletion in anucleate cells suggests RNA degradation. In nucleated cells, mRNAs are replenished by steady-state transcription. There are mechanisms that couple transcription and RNA degradation in nucleated cells (Goler-Baron et al. 2008). In platelets, expression levels of mRNAs expectedly decline in a time-dependent manner (Angenieux et al., 2016), due to RNA degradation. *Ex vivo* analysis of differential decay confirmed the depletion of linear RNA molecules and enrichment of circRNAs after 72 hrs. The reported discrepancy in number of transcripts in platelets with detectable associated protein products (Londin et al., 2014; Geiger & Burkhart 2013) is likely due to ongoing degradation of linear molecules, leaving untranslatable fragments of linear molecules. As circRNAs are enriched in platelets, the weak correlation between the transcriptome and the proteome provides further evidence that circRNAs do not contribute to the proteome.

Various studies of platelet transcriptomes showed variation in number of detectable protein-coding transcripts. These variations likely result from RNA decay and are influenced by various factors, including age, race and gender. Research has largely focused on characterizing PTES in various cells, but circRNA decay mechanisms remained unexplored. In my analysis, fluctuation in number of identified PTES transcripts is observed in one platelets sample, with only about 35% of transcripts identified in other platelets samples detectable. This large variation is due to ongoing decay of some circRNAs within platelets, as sequence read coverage across these circRNAs are not uniformly distributed or fully depleted. Correlational analysis suggest that the number of miRNA binding sites within circRNAs without uniform read coverage and their GC content may significantly influence their stability. Indeed, for some transcripts, sequence regions without read coverage directly overlap with miRcode (Jeggari et al., 2012) predicted miRNA binding sites. It is conceivable that some circRNAs without uniform read coverage are actively undergoing endonucleolytic cleavage induced by miRNAs. Although 3 circRNAs have been shown to be effective miRNA sponges (Hansen et al., 2011; Memczak et al., 2013; Zheng et al., 2016), there is evidence of miRNA induced circRNA degradation (Hansen et al., 2011). The ~65% reduction in number of identified PTES transcripts, coupled with reported high abundance of miRNAs in platelets (Landry et al. 2009; Pontes et al. 2015), plausibly suggest that circRNAs are antagonized by miRNAs and any potential miRNA sponge effect is negligible, as circRNAs are generally lowly expressed, constituting <1% of the transcriptome. Ruling out technical variability in molecular biology protocols, the reason for extensive decay within one sample is not readily clear. Without additional information about the demographics of the donors, the underlying reason can only be speculated. One plausible explanation relates to the age of platelets sampled. As platelets are

short-lived, any time difference in sample preparation may be sufficient to impact RNA integrity.

Finally, fewer PTES transcripts were identified from polyA+ platelets sample. Previous studies characterized PTES transcripts identified from polyA+ RNA fractions as linear PTES arising from trans-splicing. However, on the evidence of extensive RNA decay in platelets, reads supporting PTES in polyA+ sample likely originate from circRNAs. These transcripts were shown to be enriched for adenosine residues and were probably pulled down during isolation. It is currently unclear what impact such contamination have on expression estimates of PTES host genes and published reports of differential expression, particularly in degraded transcriptomes. It is however clear, that the vast majority of PTES transcripts are circular, at least in platelets.

5.5 Conclusion

Circular RNAs are highly stable due to their lack of free termini and resistance to exonuclease activity. In platelets, splicing and translation occur, but in the absence of nuclei, linear RNA molecules degrade in a time-dependent manner. In this chapter, I screened 4 anucleate RNAseq datasets for PTES, found anucleate samples to be enriched for PTES, subsequently adding to the growing catalog of characterized PTES transcripts. The reported weak correlation between the transcriptome and proteome of platelets is due to RNA decay and circRNA enrichment in platelets however adds to increasing evidence that PTES transcripts are non-coding. Interestingly, although additional PTES transcripts were identified from platelets, the vast majority of PTES transcripts are circular and there is no evidence of PTES transcripts arising from cytoplasmic splicing.

Chapter 6. PTES Events in development

6.1 Introduction

In the last chapter, I presented results showing that PTES transcripts are more abundant in anucleate cells, presumably accumulating due to their resistance to exonuclease digestion. Abundance of circRNAs in platelets was found to be higher than observed in samples from nucleated tissues and samples treated with RNase R to preferentially enrich circRNAs. For some genes, only reads emanating from exons involved in circRNAs were detectable. Taken together with the reported low correlation between the platelet transcriptome and proteome, these data support earlier results showing PTES transcripts as non-coding (see chapter 4). Results further raise questions about functional significance of these transcripts. In this chapter, I investigate any potential roles for PTES transcripts in either pluripotency maintenance or differentiation, by assessing their distributions in human embryonic stem cells differentiation series and screening for developmental stage-specific transcripts.

6.1.1 Epigenetic changes influence transcriptome diversity during development

Embryonic stem cells (ESC) are pluripotent, capable of self-renewing and differentiating into any somatic cell. ESC uniquely undergo both symmetrical (to produce two daughter cells of the same fate) and asymmetrical (to produce two daughter cells of different fates: pluripotent and somatic cells) cell division during differentiation (Morrison & Kimble 2006). It is thought that various signaling pathways are required to maintain the balance between both forms of cell divisions during differentiation, subsequently resulting in terminally differentiated somatic cells and resting adult stem cells (Morrison & Kimble 2006; Boland et al. 2014). Maintaining the pluripotent state of ESC involves complex transcriptional and epigenetic regulatory controls. Reports have shown that knockdown of DNA methyltransferases in mouse ESC results in hypomethylation and subsequent differentiation defects (Tsumura et al. 2006; Okano et al. 1999).

The stage-specific processing of one circRNA is already known to play a key role in development. The expression of linear canonical and circular transcripts from *Sry* during development is regulated by epigenetic modifications (Nishino et al. 2004; Nishino et al. 2011). The linear transcript is specifically expressed and translated within a developmental temporal window between days 10.5 and 12.5 in rodent embryos (Nishino et al., 2004 & 2011). Transcription of the linear and circular transcripts from *Sry* is aided by two promoters: a main promoter driving the transcription of the linear transcript and a cryptic promoter upstream of

the main promoter (Nishino et al., 2004). Both promoters are hypermethylated before the temporal window when demethyaltion of the main promoter is observed. Following the temporal window, the main promoter once again becomes hypermethylated but the cryptic promoter is demethylated, allowing for the expression of circSry in adult rodents (Capel et al., 1993; Nishino et al., 2004).

Furthermore, there is mounting evidence that the rate of of transcription elongation may affect the choice of splice sites in alternative splicing (Bentley 2014; Yue et al. 2015) and presumably PTES. During transcription, splice sites compete for spliceosomal proteins, strong splice sites outcompete weak sites and long range intron pairings are modulated by transcription elongation rates (Bentley, 2014). Studies have shown that most genes are expressed in ESC, albeit at low levels for lineage-specific transcripts, with promoters characterized by distinct chromatin signatures of either H3K4me3 for active genes, repressive H3K27me3 or both (Min et al., 2011). Modifications to these histone marks modulate expression of genes upon differentiation (Min et al. 2011; Boland et al. 2014). Moreover, increased CpG methylation within exons has been shown to improve inclusion of alternatively spliced exons (Maunakea et al. 2013; Gelfman et al. 2013; Singer et al. 2015), presumably acting as 'road blocks' that slow transcription (Schwartz et al. 2009; Choi 2010; Sati et al. 2012), long enough for splicing to complete. It is unclear whether intragenic CpG methylation plays a role in PTES directly but there are reports of correlation between alternative splicing and PTES (Zaphiropoulos 1997; Surono et al., 1999). A recent report estimated the average elongation rate of PTES genes at ~2.90 kb/min, compared to 2.29 kb/min for non-PTES producing genes (Zhang et al., 2016). In that study, HEK293 cells were transfected with 3 variants of RNA polymerase II, two of which carried mutations that either increased or decreased transcription elongation rates. Subsequently, more PTES transcripts were identified from samples transfected with the variant with fast elongation rate (Zhang et al. 2016), demonstrating a link between transcription elongation rate and PTES. As epigenetic changes evidently influence transcriptome diversity in development, assessing PTES populations during cellular differentiation may yield interesting results and deepen our understanding of mechanisms of PTES formation.

6.1.2 Non-Coding RNAs (ncRNAs) in pluripotency maintenance and differentiation

In addition to epigenetic modifications, a core set of transcription factors (including *NANOG, SOX2 & OCT4*) have been shown to be integral to regulatory mechanisms necessary for pluripotency maintenance, by directly promoting the expression of ESC-specific genes and indirectly suppressing cell lineage commitment genes. Many ncRNAs are now understood to aid ESC pluripotency maintenance by recruiting or tethering histone modifiers and

methyltransferases to direct chromatin modification of genes required for differentiation (Ng et al. 2012; Chen & Dent 2014; Fatica & Bozzoni 2013). Depletion of some ncRNAs in ESC can induce cell retardation (Hacisuleyman et al. 2014), suggesting unknown mechanisms of impact on pluripotency maintenance. One example is the multifunctional *FIRRE*, a long ncRNA that continues to be expressed from the inactive X chromosome (Hacisuleyman et al., 2014). Compared to wild-type mouse ESC (mESC), the growth rate of mESC lacking *Firre* was reported to be markedly reduced (Hacisuleyman et al., 2014). This retarded growth rate is accompanied by depletion of genes involved in mRNA processing and export, and inhibition of adipogenesis (Hacisuleyman et al., 2014). In mouse fibroblasts, knockdown of *Firre* resulted in loss of histone methylation (H3K27me3) on the inactivated X chromosome (Yang et al. 2015), suggesting a role in maintaining epigenetic state in that chromosome.

Post-transcriptionally, ncRNAs play roles in regulating the expression of both ESC-specific and development-specific transcripts (Gruber et al. 2014; Hu et al. 2012; Wang et al. 2013). A miRNA family, miR-300, has been shown to be highly expressed in ESC and down regulated upon differentiation (Hu et al., 2012). These miRNAs are understood to aid the suppression of transcripts required for lineage commitment, subsequently helping maintain the pluripotent state (Hu et al., 2012). Conversely, some miRNAs (including miR-145) target the core transcription factors, thus play roles in development (Wang et al., 2013). Some long non-coding RNAs (ncRNAs with size > 200bp) have been shown to act as competing endogenous RNAs, competing for miRNA binding and acting as decoys. LincRNA-RoR, for instance, is a long intergenic ncRNA with binding sites for miR-145 and acts as a potent decoy, reducing their effect on expression levels of ESC-specific transcripts (Cheng & Lin 2013; Wang et al. 2013).

6.1.3 PTES transcripts do not have uniformly ascribed functional significance

To date circRNAs with ascribed functions act as cytoplasmic miRNA sponges or nuclear regulators of transcription. Three circRNAs from *CDR1* antisense (CDR1as [Hansen et al., 2011; Memczak et al., 2013]), *Sry* (Hansen et al., 2011) and *HIPK3* (Zheng et al., 2016) loci have recently been shown to be potent miRNA sponges, sequestering miRNAs from target linear transcripts. Memczak et al., (2013) reported reduction in mid-brain size of zebrafish upon transfection of minigene constructs expressing CDR1as circRNA. This circRNA harbours over 70 miRNA binding sites and effectively sequesters miR-7, resulting in the observed reduction in mid-brain size. As this investigation involved the use of minigenes expressing circRNA, it is not clear whether endogenous levels of circRNAs are equally impactful and to what extent other circRNAs harbour miRNA binding sites. Two circRNAs (comprising of exons and introns

- EIcircRNAs [Li et al., 2015]) were recently reported to associate with snRNAs and RNA Polymerase II, and may facilitate the transcription of their respective parental genes.

However, there has also been a report of a linear PTES from a lncRNA which acts to maintain ESC pluripotency. Wu et al., (2013) reported the identification of 4 linear PTES transcripts (*CSNK1G3, ARHGAP5, FAT1* and *RMST*) with expression patterns suggestive of roles in pluripotency maintenance and differentiation. One from *RMST* locus recruits polycomb repressive complex 2 - PRC2, affecting the expression of development-specific transcripts and promoting pluripotency maintenance (Wu et al. 2013). By RNA immunoprecipitation assays, the authors demonstrated that the PTES from RMST interacts with *NANOG* and *SUZ12*, a component of PRC2. Knockdown of this PTES resulted in decrease in expression levels of core transcription factors necessary for pluripotency maintenance (Wu et al., 2013). The mechanism by which this PTES regulates these core transcription factors was not explored by Wu et al., (2013), but the decrease in expression levels of these ESC-specific genes was accompanied by reduction of H3K27me3 in promoters of *GATA4*, *GATA6* and *PAX6*, key lineage-specific transcription factors, and subsequent up regulation of these genes (Wu et al., 2013).

Although this finding raises the possibility of identifying other PTES transcripts with similar roles, identification of PTES transcripts was from a H9 sample with ~1 Million reads, albeit long (~350bp) reads. Identified transcripts were supported by single reads and in the absence of biological replicates, no statistical analysis was performed. Additionally, their attempt to experimentally verify the structure of this PTES using RT-PCR was inconclusive as no positive controls were used.

6.2 Aims

To address some of these limitations and to further identify PTES transcripts with expression suggestive of roles in pluripotency maintenance and differentiation, collaboration with Prof. Lako (Newcastle University, UK) was established. Prof. Lako and colleagues are investigating the effect of insulin growth factor 1 (IGF-1) on differentiation of human ESC (HESC) into retinal cells. RNA extracts from H9 ESC at 3 time points (days 0, 45 & 90) were sequenced, along with RNA from samples treated with IGF-1 to facilitate differentiation. IGF-1 treatment has been shown to increase the rate of developing retinal photoreceptors from differentiation of H9 by up to 40% (Mellough et al. 2015). It is also reported to facilitate the differentiation of mesenchymal cells into nueronal-progenitor cells (Huat et al. 2014), further suggesting that any circRNAs with roles in differentiation may be affected by IGF-1 treatment.

In this chapter, my specific aims were to:

- Identify PTES transcripts in H9 ESC differentiation series:
    - Explore properties of PTES transcripts upon differentiation
    - Identify factors that may affect PTES abundance during differentiation
- Investigate changes in expression of PTES and their linear counterparts during cellular differentiation.
- Identify PTES transcripts with expression patterns similar to that of ESC-specific genes and suggestive of roles in pluripotency maintenance.
- Assess any transcriptome-wide effect of IGF-1 treatment on PTES biogenesis and abundance

6.3 Results

To elucidate potential functional relevance of PTES, H9 ESC were differentiated into retinal cells and RNA extracts in triplicates from 3 time points (days 0, 45 and 90) (Mellough et al. 2012; Mellough et al. 2015) and were sequenced by AROS (Arhus, Sweden, see 2.5.1). In parallel, RNA extracts from H9 ESC cells treated with IGF-1 to facilitate differentiation and sequenced. From each sample, 100bp pair-end reads were generated, resulting in library sizes of over 100 million reads. According to findings reported by the ENCODE consortium (https://genome.ucsc.edu/ENCODE/protocols/dataStandards/ENCODE_RNAseq_Standards_V1.0.pdf), this sequencing depth is required for identification of novel transcripts. For each sample, reads were merged, quality controlled and screened for PTES, using methods detailed in chapter 2. PTESFinder analyses of these samples subsequently identified 58,794 distinct PTES transcripts, from 8729 genes, resulting in ~7 PTES transcripts per gene and ~5 PTES supporting reads per PTES transcript (Table 6.1A-D). Notably, PTES producing genes represent ~20.4% of GENCODE annotated genes (n=42,785). On average, ~5X more PTES supporting reads were identified from differentiated time points than in day 0 (Table 6.1B). In contrast, ~1.5X more canonical junction reads were observed in undifferentiated (day 0) samples than in higher time points. However, more canonical junctions are observed in differentiated (days 45 & 90) time points. The highest number of identified PTES transcripts, PTES supporting reads and number of PTES producing genes were observed in day 45 (Table 6.1B & D). When treatment groups are compared, there is no apparent difference in the number of PTES transcripts identified and abundance between treatment groups (Table 6.1C, Fig. 6.1), suggesting that IGF-1 treatment has no global effect on PTES. This observation is consistent with a recent study, where another growth factor (epidermal growth factor [EGF]) was found to have little or no effect on circRNA abundance in multipotent cells, concluding that PTES transcripts may not play roles in signaling cascades that lead to lineage commitment(Enuka et al. 2015).

A.

ID	PTES transcripts	PTES Reads	PTES Genes	Canonical Junctions	Canonical Junction Reads	Library Size	Genome Reads (%)	Transcriptome Reads (%)	Average Reads per PTES	Average PTES per Gene
day_zero_a	6101	16051	2787	158362	8551592	124825004	82.21	75.71	2.63	2.19
day_zero_b	6098	16184	2885	169452	10742338	142274904	91.02	82.36	2.65	2.11
day_zero_c	7256	19345	3332	165385	12714653	140196638	90.97	89.12	2.67	2.18
45_control_a	20307	116880	5312	165701	5188215	110773232	83.59	75.55	5.76	3.82
45_control_b	18159	96388	5345	182550	8397547	138387142	92.51	88.91	5.31	3.40
45_control_c	18537	95785	5226	180737	7889224	146551576	93.13	88	5.17	3.55
45_IGF_1a	14098	65171	4501	161901	6162224	105685200	83.04	80.05	4.62	3.13
45_IGF_1b	18760	99849	5333	182789	8834255	133788724	92.14	89.37	5.32	3.52
45_IGF_1c	17331	88330	5126	179547	8632502	140203130	92.70	88.05	5.10	3.38
90_control_a	11858	45983	4245	165512	4204966	116522226	82.19	51.94	3.88	2.79
90_control_b	16599	79617	4942	187922	9631364	160607052	91.34	76.41	4.80	3.36
90_control_c	18398	100061	5283	178941	6954842	153237218	92.66	72.44	5.44	3.48
90_IGF_1a	17641	94575	5132	178608	7371972	132360610	92.03	78.89	5.36	3.44
90_IGF_1b	16498	78816	4935	188803	9584313	155002132	91.08	75.28	4.78	3.34
90_IGF_1c	17156	90151	5119	182391	8818808	145942028	92.10	87.21	5.25	3.35

B. Mean summary (by time point)

Time Point	PTES transcripts	PTES Reads	PTES Genes	Canonical Junctions	Canonical Junction Reads	Library Size	Genome Reads (%)	Transcriptome Reads (%)	Average Reads per PTES	Average PTES per Gene
0	6485	17193	3001	164400	10669528	135765515	88.07	82.40	2.65	2.16
45	17865	93734	5141	175538	7517328	129231501	89.52	84.99	5.21	3.47
90	16358	81534	4943	180363	7761044	143945211	90.23	73.70	4.92	3.29

C. Mean summary (by treatment)

Group	PTES transcripts	PTES Reads	PTES Genes	Canonical Junctions	Canonical Junction Reads	Library Size	Genome Reads (%)	Transcriptome Reads (%)	Average Reads per PTES	Average PTES per Gene
control	17310	89119	5059	176894	7044360	137679741	89.24	75.54	5.06	3.40
igf	16914	86149	5024	179007	8234012	135496971	90.52	83.14	5.07	3.36
zero	6485	17193	3001	164400	10669528	135765515	88.07	82.40	2.65	2.16

D. Mean summary (by time point & treatment)

Time Point	Group	PTES transcripts	PTES Reads	PTES Genes	Canonical Junctions	Canonical Junction Reads	Library Size	Genome Reads (%)	Transcriptome Reads (%)	Average Reads per PTES	Average PTES per Gene
45	control	19001	103018	5294	176329	7158329	131903983	89.74	84.15	5.41	3.59
90	control	15618	75220	4823	177458	6930391	143455499	88.73	66.93	4.70	3.21
45	igf	16730	84450	4987	174746	7876327	126559018	89.29	85.82	5.01	3.34
90	igf	17098	87847	5062	183267	8591698	144434923	91.74	80.46	5.13	3.38
0	zero	6485	17193	3001	164400	10669528	135765515	88.07	82.40	2.65	2.16

Table 6.1. **Summary of PTES identified from H9 ESC differentiation series.** A) Summarised by sample B) by time point C) by treatment group and D) by time point and treatment group.

Of the 45,311 distinct PTES transcripts identified from control samples, only 10,437 transcripts were identified from both differentiated (days 45 & 90) and undifferentiated (day 0) samples. Over thirty-two thousand transcripts were uniquely identified from differentiated samples and only 2790 transcripts are unique to day 0 (Fig 6.1A). For downstream analysis, read counts for both PTES and canonical junctions were summed across biological replicates. Clustering by Euclidean distance between samples expectedly shows day 0 clustered away from other samples, further highlighting the difference in PTES populations between stages (Fig. 6.1B). PTES counts separates samples by time point (and not treatment group) (Fig 6.1B-C), underscoring the limited effect of IGF-1 treatment on PTES. Pairwise comparisons of PTES from each sample shows higher correlation coefficients (min: 0.78) between differentiated samples, than with day 0 samples (max: 0.59) (Fig. 6.2).

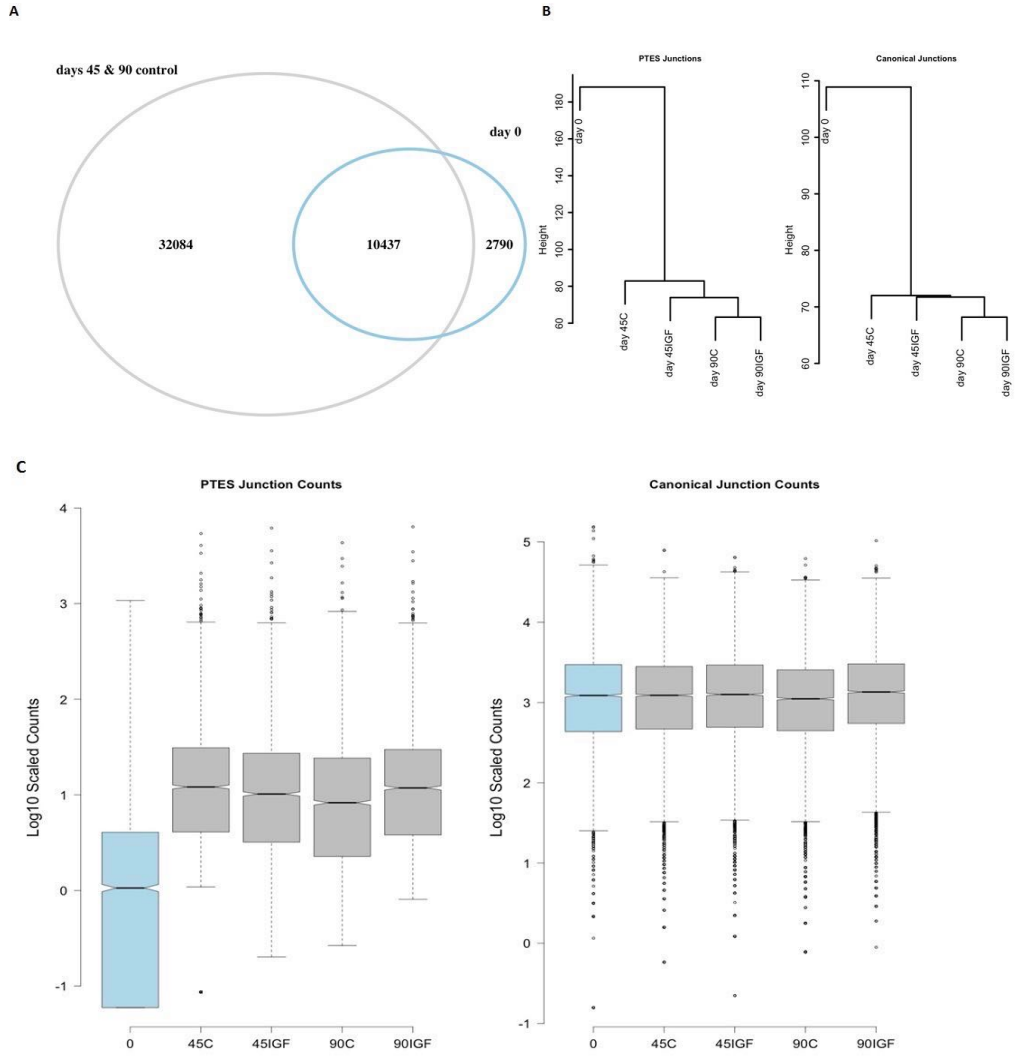


Figure 6.1. **Exploratory analysis of PTES in HESC differentiation.** A) Number of PTES transcripts identified from differentiation series (days 0, 45 & 90), without IGF-1 treatment B) dendrogram showing hierarchical clustering of samples using PTES and canonical junction counts C) Boxplots of Log10 normalized PTES and canonical junction counts. More PTES supporting reads are identified in differentiated samples that in day 0; contrasting with slightly higher number of canonical junction reads in day 0, relative to higher time points.

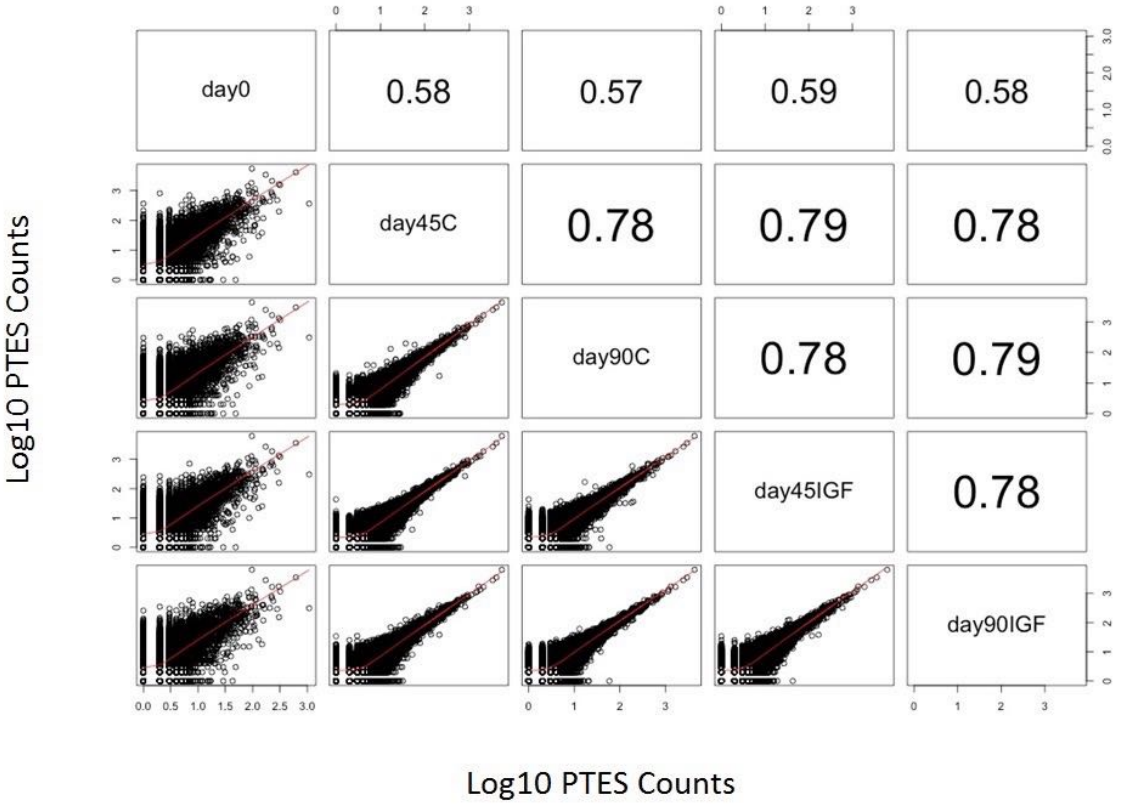


Figure 6.2. **Correlational analysis of PTES in HESC differentiation.** Pairwise scatterplots and correlation of PTES abundance across time points. Correlation coefficients are higher for comparisons between days 45 & 90, than comparisons between day 0 and differentiated (days 45 & 90) samples.

6.3.1 Differences in PTES abundance correlates with expression levels of RNA Binding Proteins (RBP)

To identify factors affecting the number of PTES transcripts identified between differentiation stages, I first estimated the expression of all GENCODE annotated transcripts. Three RNA binding proteins (RBP) have been shown to have roles in PTES biogenesis. Two splice factors, *MBNL* (Ashwal-Fluss et al., 2014) and *QKI* (Conn et al. 2015), promote secondary structures favourable to PTES biogenesis, while *ADAR*, an RNA editing enzyme inhibits the formation of dsRNA and PTES (Rybak-Wolf et al. 2015). For these reasons, I profiled the expression of these RBPs in all 3 time points, reasoning that the striking difference in number of identified transcripts may be due to expression levels of these proteins. As controls, I also profiled the expression of *NANOG*, an ESC-specific transcription factor and *MITF*, a transcription factor shown to be differentially expressed during differentiation of HESC into retinal cells (Liao et al. 2010).

MBNL expression is slightly elevated in differentiated samples (p-value: 0.0002), but *QKI* levels are similar across time points (Fig. 6.3A), suggesting that the observed increase in PTES

biogenesis may not solely be attributed to these 2 splice factors. However, there is ~3X reduction in expression levels of *ADAR* (p-value: 5.496 x 10^{-10}) between differentiated samples and undifferentiated samples (Fig. 6.3B), suggesting a role for *ADAR* in suppressing PTES biogenesis in pluripotent cells. This observation supports a recent report of increase in PTES events in mouse P19 embryonic carcinoma cells following *ADAR* knockdown (Rybak-Wolf et al., 2015).

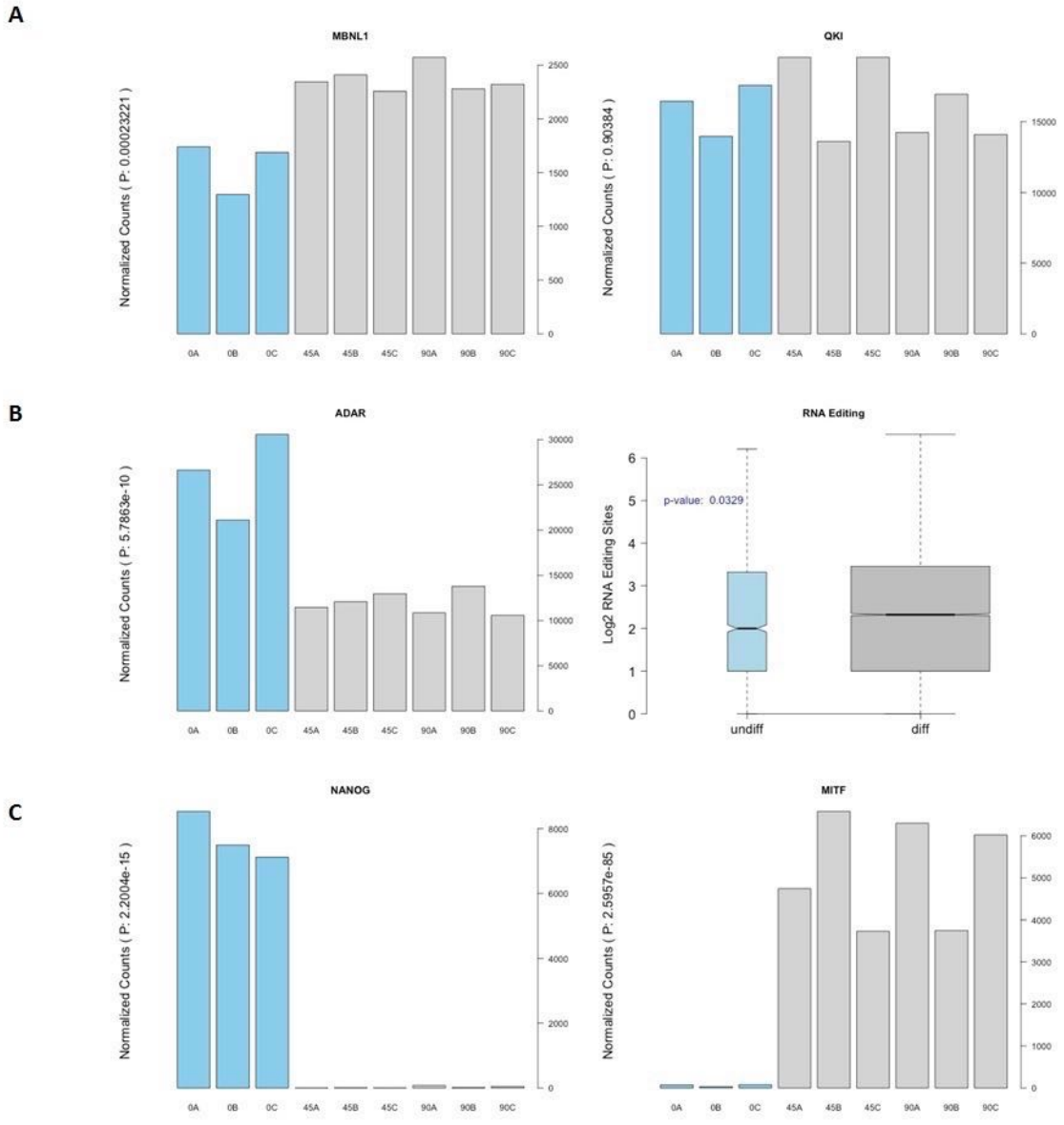


Figure 6.3. **RNA Binding proteins affect PTES biogenesis & abundance. A)** Expression estimates of two splice factors, *MBNL* and *QKI*, and **B, left)** RNA editing enzyme, *ADAR*, upon HESC differentiation. Right) Box plots showing number of RNA editing sites within 1000bp of PTES flanking introns. **C)** Expression estimates of transcription factors, *NANOG* & *MITF*, shown here as controls.

It is conceivable that PTES transcripts uniquely observed in differentiated samples are characterized by high RNA editing sites in flanking introns. To assess this, I extracted the genomic positions of the first 1000bp intronic sequence proximal to PTES junctions identified from both stages (differentiated and undifferentiated). I then compared the genomic positions of these flanking introns to genomic positions of RNA editing sites available from rnaedit.com (Chen 2013; Porath et al., 2014). Approximately 52.7% of PTES transcripts identified from undifferentiated samples have at least one RNA editing site within 1000bp of flanking introns, with a median of 4 editing sites (Fig. 6.3B). Conversely, ~56% of transcripts identified in differentiated samples have editing sites and a median of 5 RNA editing sites, higher than observed for transcripts in undifferentiated samples (p-value: 0.03, Wilcoxon rank test), consistent with some of these transcripts being suppressed in day 0 due to A-to-I editing weakening intron pairing and subsequent PTES formation. An example may be a PTES transcript from *DHDDS* locus between exons 5 and 6 (DHDDS.6.5). This structure is only observed in differentiated time points (highest in day 90) and flanking introns have 12 and 27 RNA editing sites respectively (Fig 6.4). This gene is implicated in retinal disorders (Zuchner et al. 2011) and expressed at comparable levels in all samples, thus, ruling out transcriptional silencing as reason for not detecting this PTES transcript in day 0.

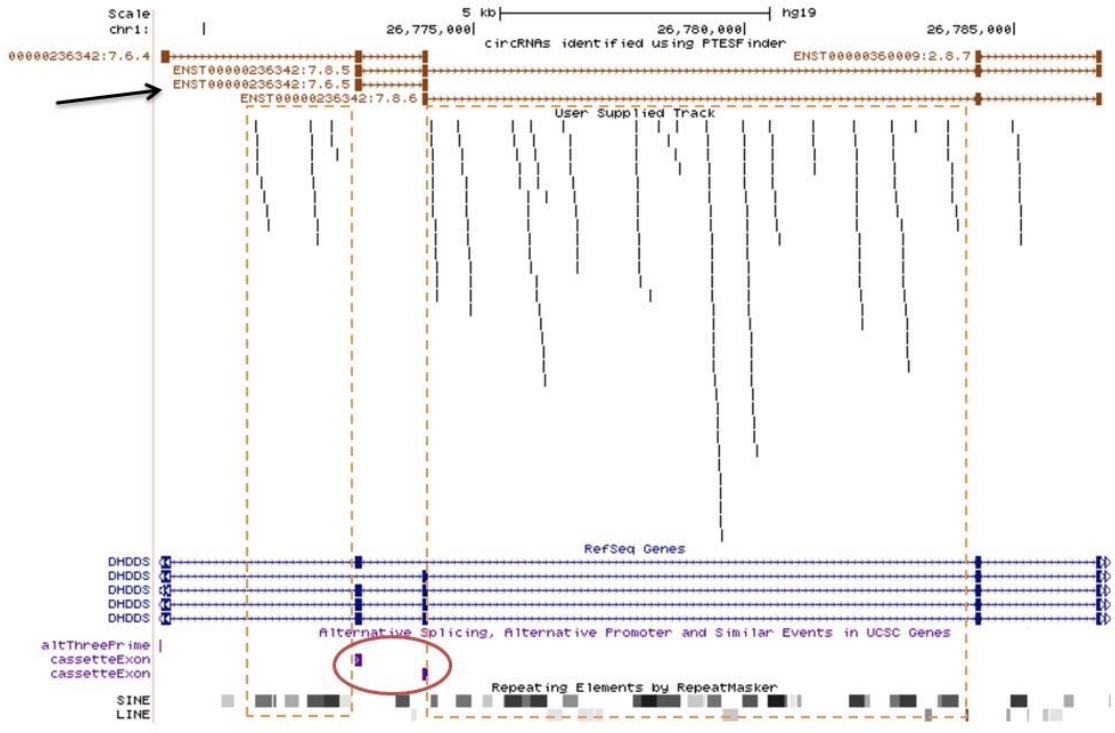


Figure 6.4. **RNA editing sites flanking DHDDS.6.5**. RNA editing sites (yellow boxes) within flanking introns of DHDDS.6.5, a PTES observed only in differentiated samples. Backspaced exons (6 & 5) are known cassette exons (red highlight) and are flanked by numerous *Alu* repeats.

As some transcripts observed in day 0 have known RNA editing sites and are not suppressed, *ADAR* activity may not be the sole reason for PTES suppression in this time point. Further profiling of expression patterns of endoribonucleases shows their expression to be higher or comparable in day 0 samples relative to other time points (Fig. 6.5). Increases in expression of endoribonucleases have not been shown to affect PTES abundance directly; however, as circRNAs are resistant to exonucleases, a decay pathway may conceivably include linearisation of circRNAs by endoribonucleases. The high abundance of some profiled endoribonucleases in day 0 may contribute to the suppression of PTES abundance in this time point. *DIS3*, a known component of the exosome with reported endoribonucleolytic activity (Tomecki & Dziembowski 2010), and *ZC3H12A* are particularly elevated in undifferentiated samples (p-values: 1.2×10^{-05} and 4.4×10^{-05}, respectively), relative to other time points.

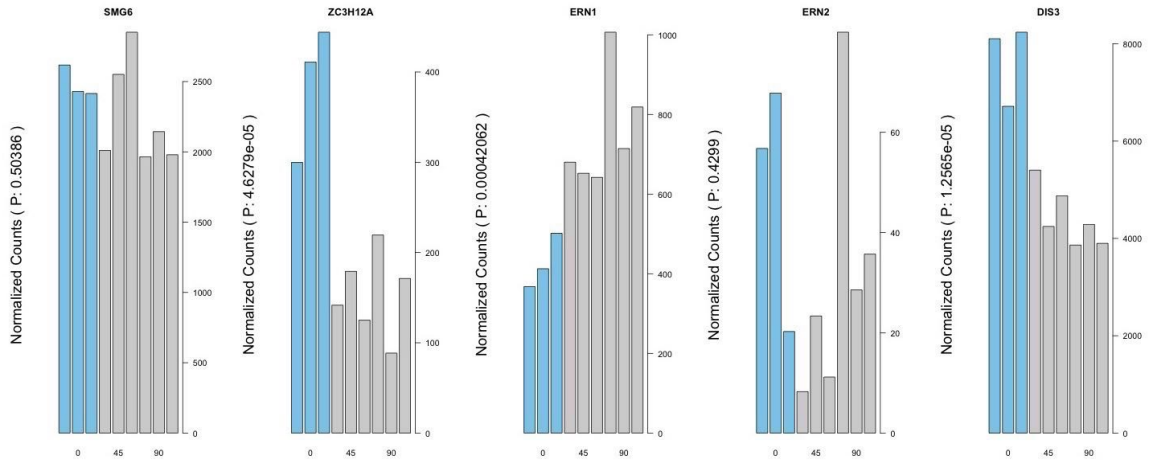


Figure 6.5. **Expression profiles of endoribonucleases**. Expression estimates of endoribonucleases are higher or comparable in ESC, relative to other time points. Expression estimates of these genes were profiled in undifferentiated (blue) and differentiated(grey) samples.

6.3.2 Properties of PTES transcripts vary between developmental stages

To assess the potential link between intragenic DNA methylation and PTES, I first compared the unspliced sizes of the longest PTES transcripts identified from each locus, in both undifferentiated and pooled differentiated samples. As DNA methylation impacts the rate of transcription elongation and subsequent formation of secondary structures favourable for PTES, I reasoned that transcript sizes may vary upon differentiation and correlate with intragenic methylation levels. Indeed, a noticeable size variation was observed. Transcripts identified from higher time points are significantly longer than transcripts in undifferentiated time points (p-value < 9.1×10^{-39}, Wilcoxon rank test)(Fig 6.6). Similarly, the number of exons included within inferred PTES transcripts is significantly higher for transcripts identified in differentiated samples than transcripts identified in day 0 (p-value < 9.1×10^{-39}, Wilcoxon rank

test). The observed size and exon count variation may be due to an increase in transcription elongation rates upon differentiation.

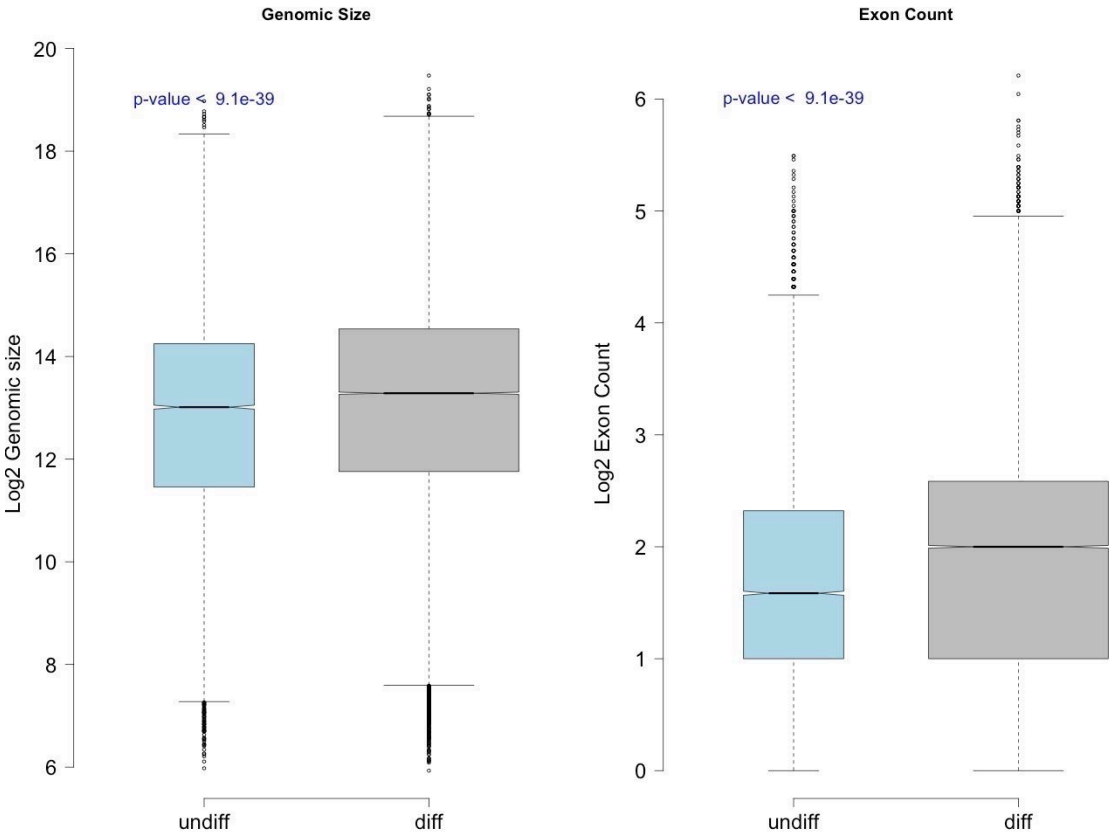


Figure 6.6. **PTES size and Exon count variation across differentiation time points**. Genomic sizes and number of exons included within circRNAs identified from undifferentiated and differentiated samples. Box widths reflect variation in number of data points within each group.

Profiling the expression of both PTES exons and non-PTES exons, there is a striking difference in profiles when day 0 is compared to day 45, relative to comparisons between day 45 and day 90. In the comparison between day 0 and day 45 (Fig 6.7, left panel), PTES exons appear to be generally expressed at low levels in day 0 relative to day 45. The slopes are noticeably different for both groups of exons (PTES and non-PTES), and the derived correlation coefficient is lower for PTES exons (than for non-PTES exons) in the first comparison but higher in comparisons between days 45 & 90 (Fig 6.7, right panel). These observations indicate global changes in regulation of PTES exons during early differentiation.

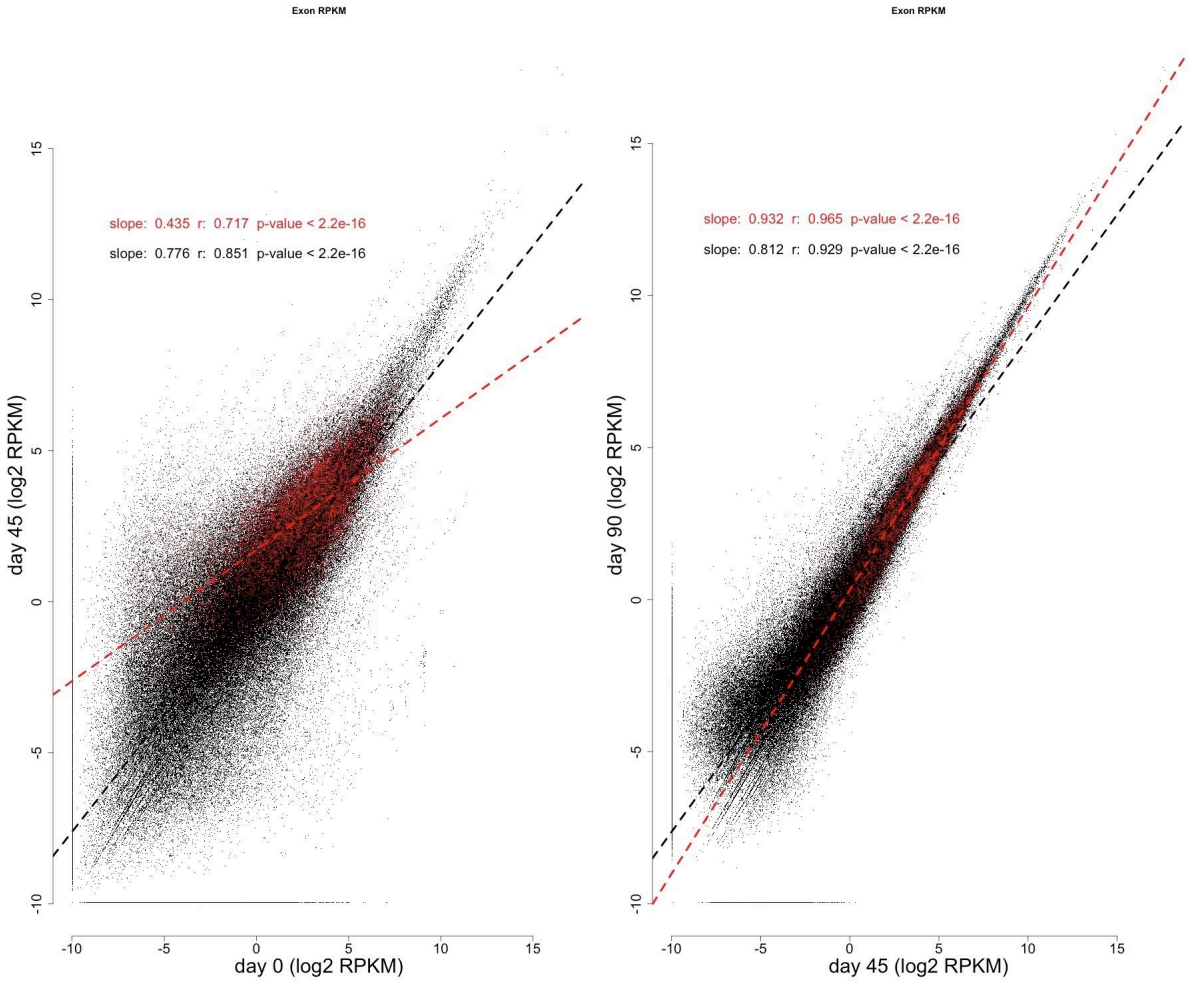


Figure 6.7. **Expression profiles of PTES and non-PTES exons** Expression estimates of PTES and non-PTES exons between time-points. Regression lines show distinct expression profiles between PTES exons (red) and non-PTES exons (black). The slopes, correlation coefficients and p-values are shown for each comparison.

Patterns of CpG methylation may contribute to these observations and the reduced PTES abundance in ESC, either by transcriptional silencing or by regulating transcription elongation rates. To assess the level of CpG methylation of promoter regions, I compared the genomic coordinates of 1000bp proximal to transcription start sites to CpG methylation data published in Liu et al., (2014). Using bisulfite sequencing, CpG methylation sites were identified from various time points following H9 ESC differentiation into retinal pigment epithelium (Liu et al., 2014). Results show similar medians of CpG methylation between PTES producing genes and non-PTES producing genes across time points (Fig. 6.8A), suggesting that promoter CpG hypermethylation may not be a significant reason for reduced PTES biogenesis in day 0. However, there is progressive reduction in intragenic CpG methylation upon differentiation (Fig 6.8B). Median CpG methylation level is higher in PTES exons than in non-PTES exons in day 0. This is however, reversed upon differentiation, suggesting that CpG demythylation occurs more frequently in PTES exons.

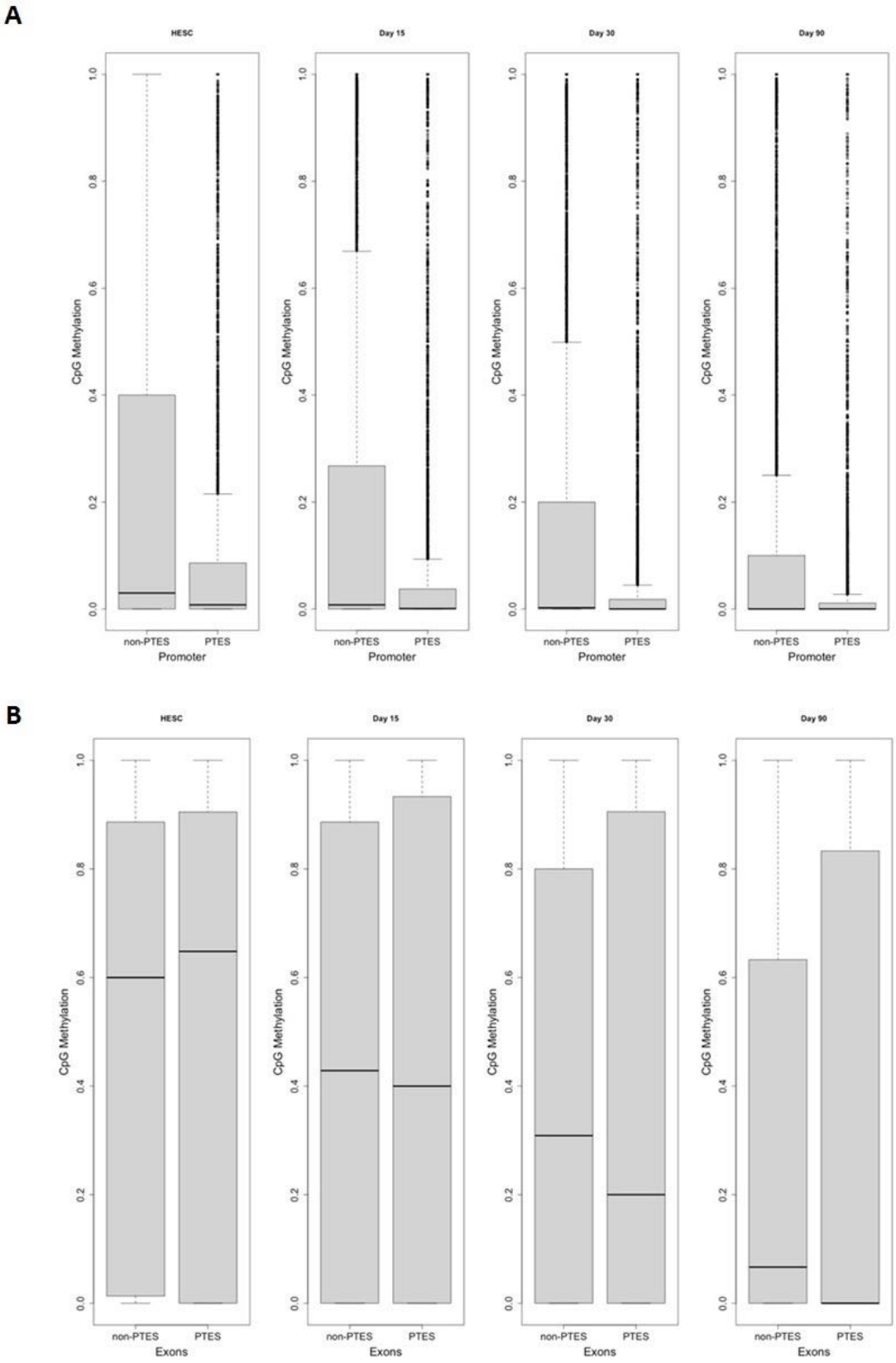


Figure 6.8. **CpG methylation profiling of PTES and non-PTES producing genes.** A) CpG methylation levels in 1000bp proximal to promoters and B) exonic regions of PTES and non-PTES producing genes.

6.3.3 PTES transcripts include exons skipped during alternative splicing

The observed discordance between numbers of canonical junction reads (which are higher in undifferentiated samples) and number of distinct canonical junctions (higher in differentiated samples) may be due to alternative splicing. Various reports posit that mechanistically PTES formation may involve re-splicing of exons within discarded lariat intermediates (Salzman et al. 2012; Jeck et al. 2013; Barrett et al. 2015), raising the likelihood that PTES transcripts comprise of skipped exons. To assess this, I inferred skipped exons from canonical junctions where the splice junction is between non-consecutive exons. For instance, a canonical junction between exons 5 & 7, suggest skipping of exon 6. Results show that PTES transcripts significantly consist of skipped exons (p-value < 2.2×10^{-16}, Fisher's exact test). Of 374,237 distinct exons within GENCODE v. 19, 74,198 are PTES exons and 33,315 are identified as skipped exons from canonical junctions. PTES exons and skipped exons overlap by 17,099 exons, 11,469 of which are UCSC knowngene annotated cassette exons. Notably, DHDDS.6.5 (example used above) consists of two cassette exons that are not detected as skipped in day 0 samples, further bolstering the link between exon skipping and biogenesis of that transcript in differentiated time points. These results, however, suggest that many constitutively spliced exons are also subjected to PTES.

6.3.4 Change in PTES abundance correlates with change in canonical junction expression

Of interest is the change in PTES abundance between time points, in order to identify transcripts likely to have functional relevance. To that end, I first normalized PTES and canonical junction counts with total observed junction counts per sample. I then compared ratios of PTES abundance between time points to equivalent ratios derived using total canonical junction counts from respective host genes. Results in Fig. 6.9A-B show that, for the majority of PTES transcripts, the observed changes between time points tracks with changes in the overall transcription output of host genes. Strikingly, regression lines and derived slopes suggest that the change in PTES abundance increases upon differentiation. For instance, comparing changes in PTES abundance between days 0 & 45 to that of canonical junctions, the derived slope is ~0.5, suggesting that a 2X increase in canonical junction expression from day 0 to day 45 correlate with ~4X increase in PTES abundance. This rate marginally reduces when expression changes in higher time points are considered, where the slope is ~0.7, suggesting that a 2X increase in canonical junction expression between day 45 and day 90 correlates with ~3X increase in PTES abundance. A plausible explanation for this relationship may be that

expression of PTES transcripts is generally not as tightly regulated as that of linear transcripts and their stability results in accumulation.

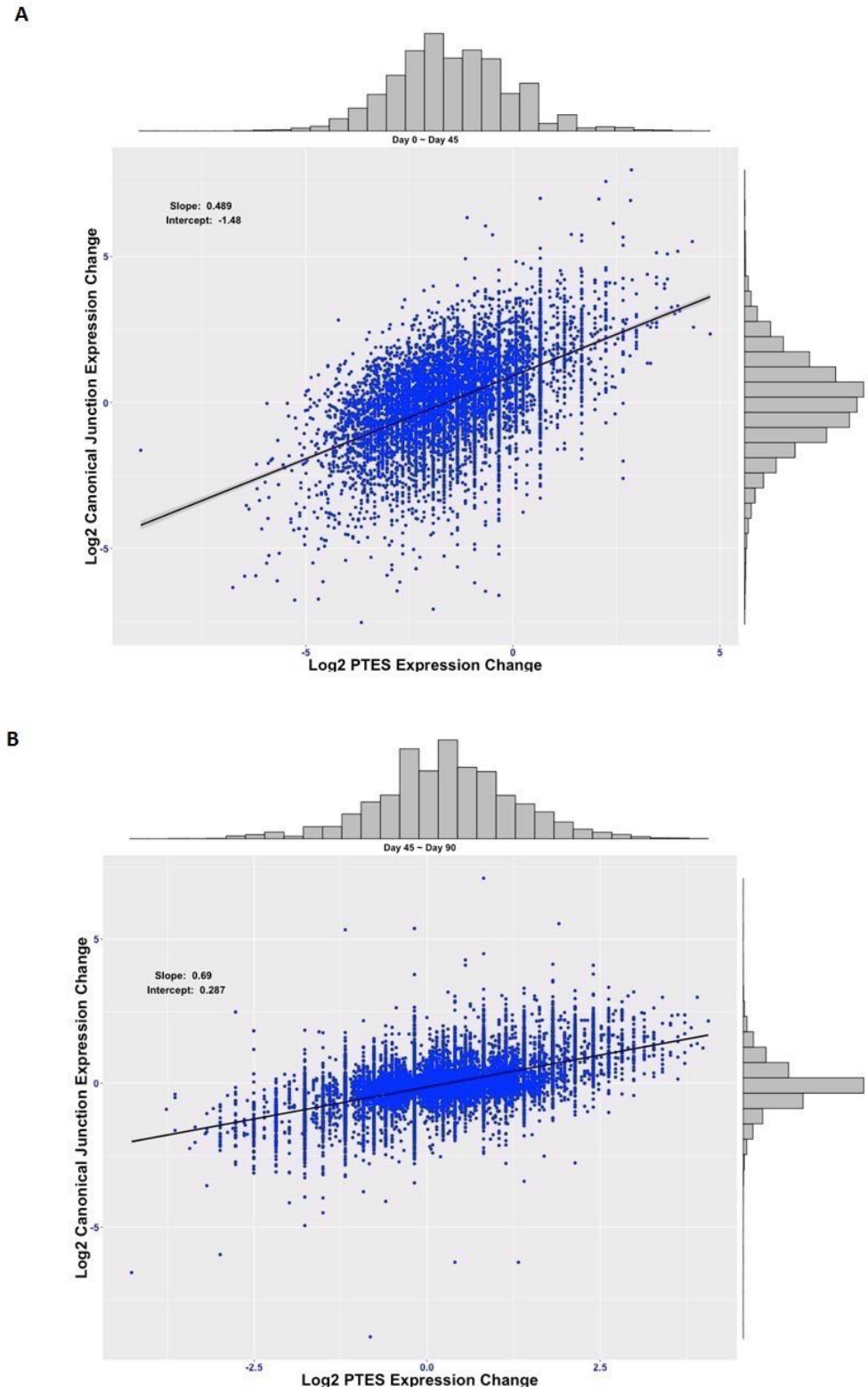


Figure 6.9. **Change in Canonical junction and PTES expression across time-points.** Change in total canonical junction expression compared to change in PTES expression between A) Day 0 and Day 45C, and B) Day 45C and Day 90C. Each data point represents the ratio of PTES abundance between time points compared and that of canonical junctions between the same time points.

To identify underlying trends in expression patterns of transcripts from PTES producing genes across time points, I performed cluster analysis of summed total canonical junction counts (see methods). Observed trends show that ~27% of all genes (n = 20,064) have decreasing expression levels of canonical junctions upon differentiation and may be ESC-specific (clusters 4 and 5 in top panel of Figure 6.10). Furthermore, <7% of all genes show marked down regulation of canonical junction expression in differentiated samples relative to undifferentiated samples (cluster 4, top panel, Fig 6.10). It is also striking that, with the exception of clusters 4 and 5, the majority of PTES transcripts associated with each cluster, have expression patterns similar to the overall canonical junction abundance of host genes (bottom panel, Fig 6.10).

Despite most PTES transcripts appearing to track canonical junction expression, some PTES transcripts do exhibit expression patterns that deviate from those of cognate canonical transcripts (Fig. 6.10). This population of transcripts likely includes transcripts with low counts - perhaps false positives. However, they may also include transcripts with expression patterns independent of cognate linear transcripts and of functional relevance.

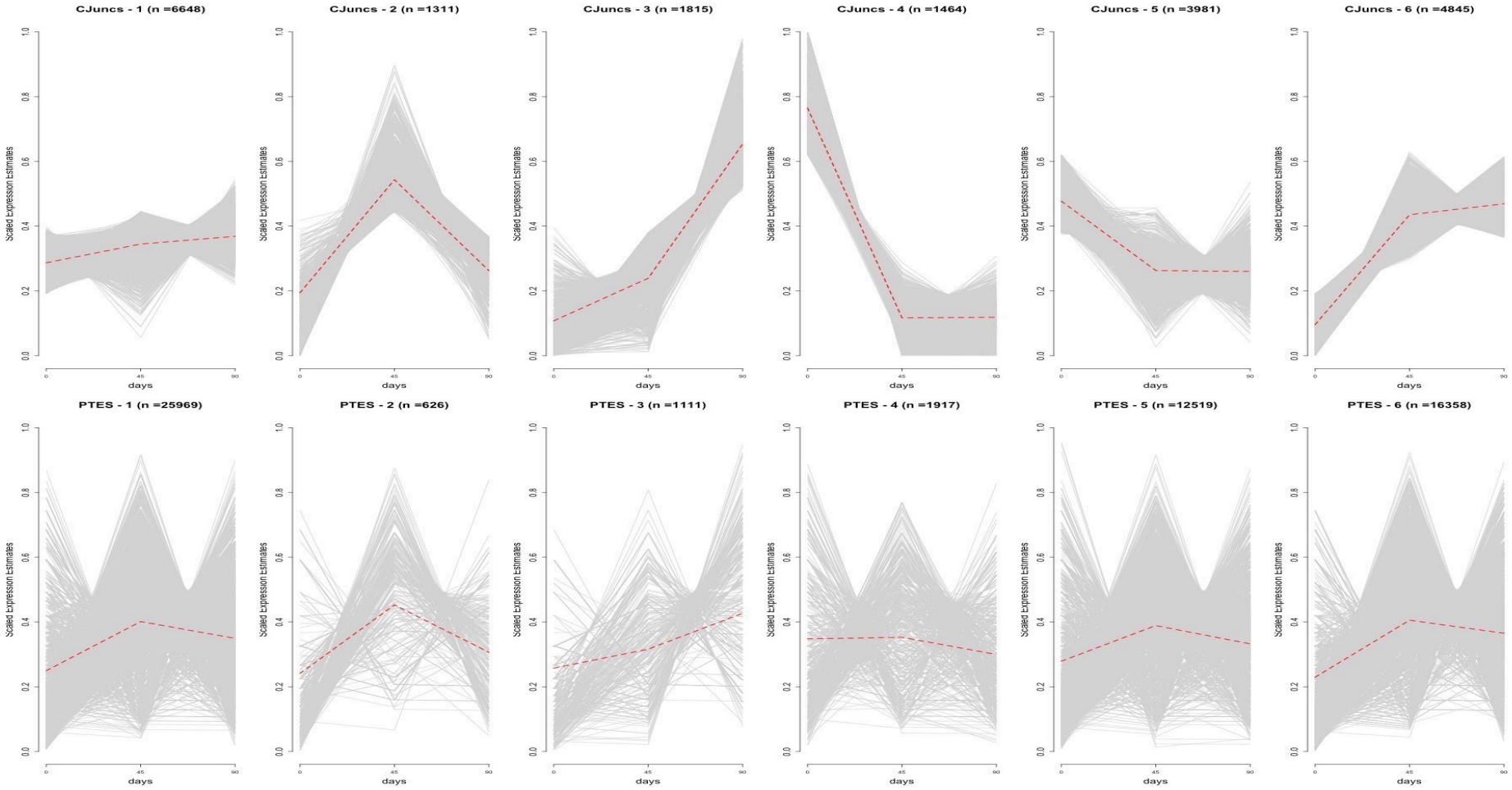


Figure 6.10. **KMeans clusters of PTES and canonical junction counts.** Six clusters of total canonical junction (top panel) and PTES (bottom panel) counts across differentiation series. Clustering was performed by KMeans, using normalized expression estimates of total canonical junctions observed for each gene across differentiation series. PTES cluster assignments were based on clusters of canonical junction expression of host genes. Numbers of genes and PTES in each cluster are shown. Red lines in each plot show mean expression estimates of each cluster.

As many of the PTES transcripts identified in this study seemingly have expression profiles that track with their cognate linear transcripts, the challenge in enrichment analysis was to identify transcripts bucking this trend. A flow chart of statistical and bioinformatics analyses performed is shown in Fig 6.11. It is necessary to eliminate PTES transcripts which may show differential expression due, either to global changes in transcription levels between time points, or changes in expression levels of linear isoforms from the same locus. Given the very low PTES counts for some transcripts, it is also necessary to negate quantitative thresholding and mitigate against potential sampling bias. To achieve these, for each transcript, I first summed the PTES read counts across replicates from each time point (as above) and performed pairwise comparisons of time points using two different analyses:

1. Identification of PTES transcripts likely to be differentially expressed (DE) between two time points (A & B), controlling for global changes in gene expression between time points (sample-level DE). A contingency table for each transcript was constructed consisting of PTES junction counts in time points A and B, versus total junction counts (PTES & Canonical) observed in both samples, minus PTES counts for transcript being tested.
2. Identification of PTES transcripts differentially expressed between time points (A & B), controlling for locus specific changes in gene expression between time points (Locus-level DE). A contingency table for each transcript was constructed consisting of PTES junction counts in both time points and their associated total canonical junction counts from locus.

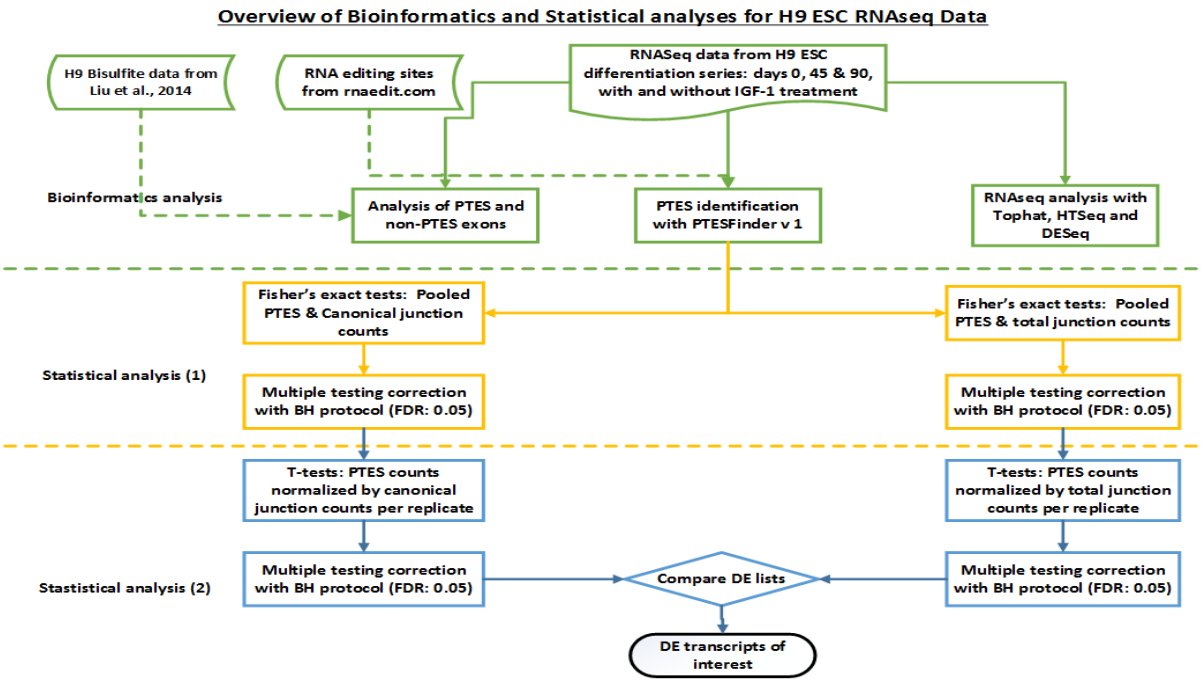


Figure 6.11: **PTES enrichment analysis workflow**. Bioinformatics and statistical analysis of H9 ESC differentiation series.

Subsequently, contingency tables were analyzed using Fisher's exact tests. Transcripts reaching significance after multiple testing correction with Benjamini-Hochberg method at false discovery rate of 0.05 were carried forward for additional analyses. These initial analyses identified a total of 8338 and 5514 potentially distinct DE transcripts respectively, across all comparisons. Fewer PTES transcripts reached statistical significance in comparisons between differentiated samples, than observed for comparisons including day 0 (Fig 6.12). Collectively, 9122 PTES transcripts reached statistical significance after both analyses and are produced from 3836 genes, representing ~44% PTES producing genes for transcripts tested.

The highest number of transcripts reaching significance after the initial analyses was observed in comparisons of day 0 and day 45C samples, with the majority enriched in day 45 in both tests. Majority of these transcripts appear in clusters 1 and 6 where expression of their host genes increases upon differentiation (Fig 6.12). Interestingly, for genes with decreasing expression during differentiation (clusters 4 and 5), more PTES transcripts reached statistical significance after locus-level analysis (bottom panel, Fig 6.12) than after sample-level analysis (top panel, Fig 6.12), a noticeable different pattern to that of other clusters. The expression profiles of the bulk of DE PTES transcripts in clusters 4 and 5 appear to increase over time or remain constant (Figure 6.12) and are different from the reductions observed for canonical junctions of their respective host genes. This observation suggests differing rates of decrease between canonical and PTES junctions upon differentiation, possibly due to differential stability, resulting in accumulation of PTES transcripts relative to linear.

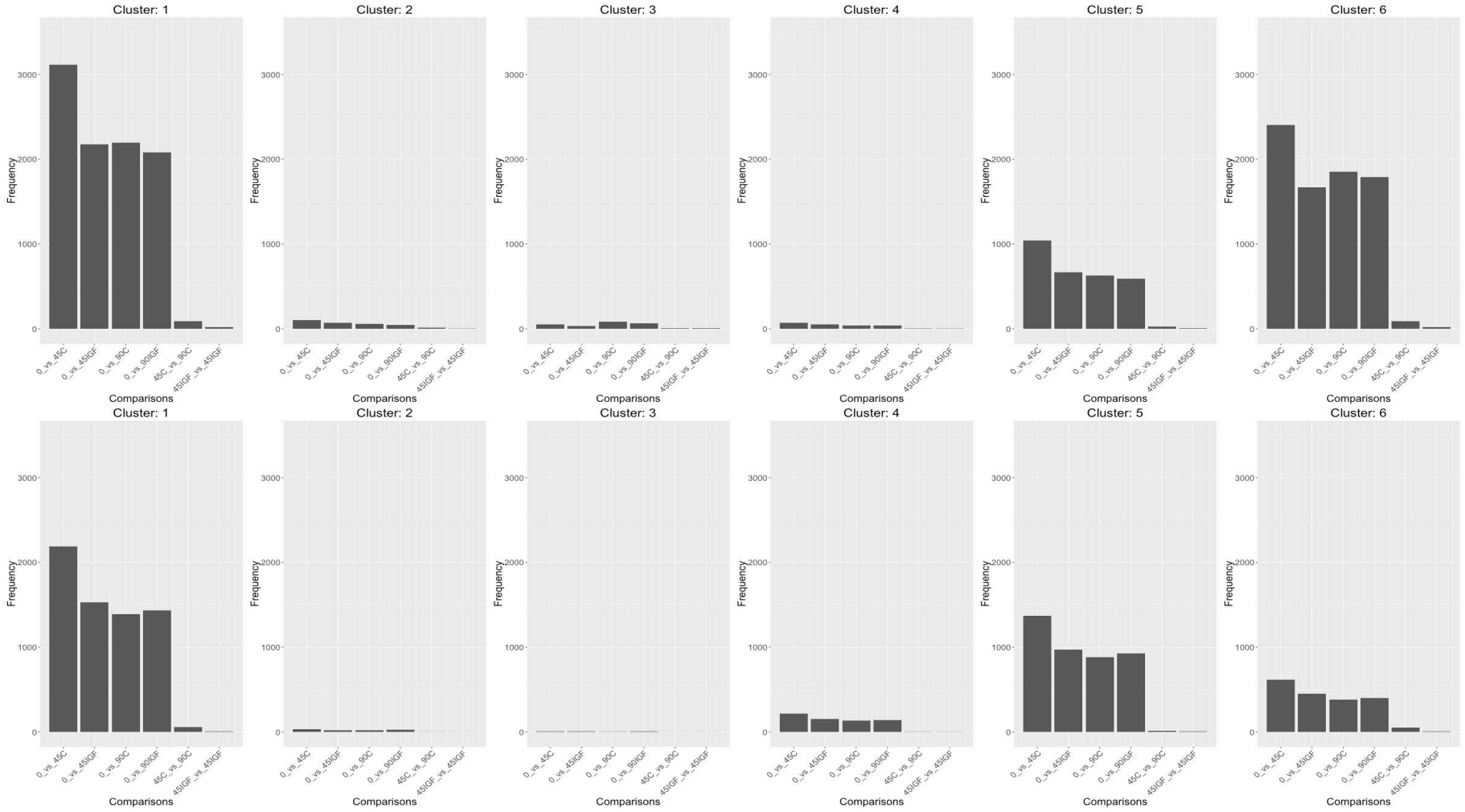


Figure 6.12. **Differentially expressed (DE) PTES transcripts before filter**. DE transcripts derived after each time point pairwise enrichment analysis, Top) after controlling for global changes in transcription and Bottom) after locus-specific changes in canonical junction expression. Transcripts are grouped in clusters to indicate expression patterns of both PTES and canonical junctions.

Having identified transcripts potentially differentially expressed at the locus specific or sample specific level, the replicate data for these transcripts were then utilized to perform more stringent two-tailed T-Tests comparing PTES abundance in biological replicates at each time point. As before, two analyses were performed, normalizing PTES counts with total canonical junction counts from respective host genes (Locus-level) for one, and total junction counts observed in each sample, for the other (Sample-level).

As we are interested in transcripts which are differentially expressed relative to both global and locus specific changes in gene expression, the DE gene lists from both analyses were compared, with the intersection of the two representing PTES transcripts differentially expressed after controlling for these global and locus specific influences. Table 5.2 summarizes the number of DE transcripts derived after all pairwise comparisons. In comparisons between differentiated samples, regardless of treatment groups, no PTES transcripts reached statistical significance. In contrast, many PTES transcripts were identified as differentially expressed in comparisons between ESC and differentiated samples.

Comparisons	Locus-level DE	Sample-level DE	Combined DE
day0_vs_day45C	60	73	7
day0_vs_day90C	89	47	13
day0_vs_day45IGF	518	542	213
day0_vs_day90IGF	156	156	38
day45C_vs_day45IGF	0	0	0
day45C_vs_day90C	0	0	0
day45IGF_vs_day90IGF	0	0	0
day90C_vs_day90IGF	0	0	0

Table 6.2. **Summary of differentially expressed (DE) transcripts.** Number of transcripts reaching statistical significance after controlling for transcription changes within host locus (Locus-level DE) and after controlling for global transcription changes between samples (Sample-level DE). DE transcripts in both lists are combined and shown in last column (Combined DE).

Strikingly, where observed, all DE transcripts are enriched in the later time point (appendix 9.5). For instance, 7 transcripts were found differentially expressed in comparison between day 0 and day 45C (Table 6.2), with all PTES transcripts enriched in day 45. Consistent with IGF-1 facilitating differentiation, the highest numbers of DE transcripts were observed after

comparison between day 0 and IGF-1 samples from higher time points (Table 6.2). But again, all PTES transcripts are enriched in the later time point.

To observe the expression patterns of both canonical and PTES junctions of transcripts reaching significance, heat maps were generated and shown in Fig 6.13. The heat maps in Fig 6.13A show variation in canonical junction (left) and PTES (right) expressions for all DE transcripts across time points after comparisons using samples not treated with IGF1. It is noticeable that the expression profiles of canonical junctions are seemingly uniform across time points, with a small number of genes having higher expression in day 0 samples. This is noticeably different from the profiles of DE PTES transcripts from the same genes, as virtually all transcripts show increased expression in differentiated samples (days 45 & 90), consistent with accumulation. This pattern is even more striking for the larger number of DE transcripts observed in comparisons between day 0 and IGF-1 samples from higher time points (Fig 6.13B).

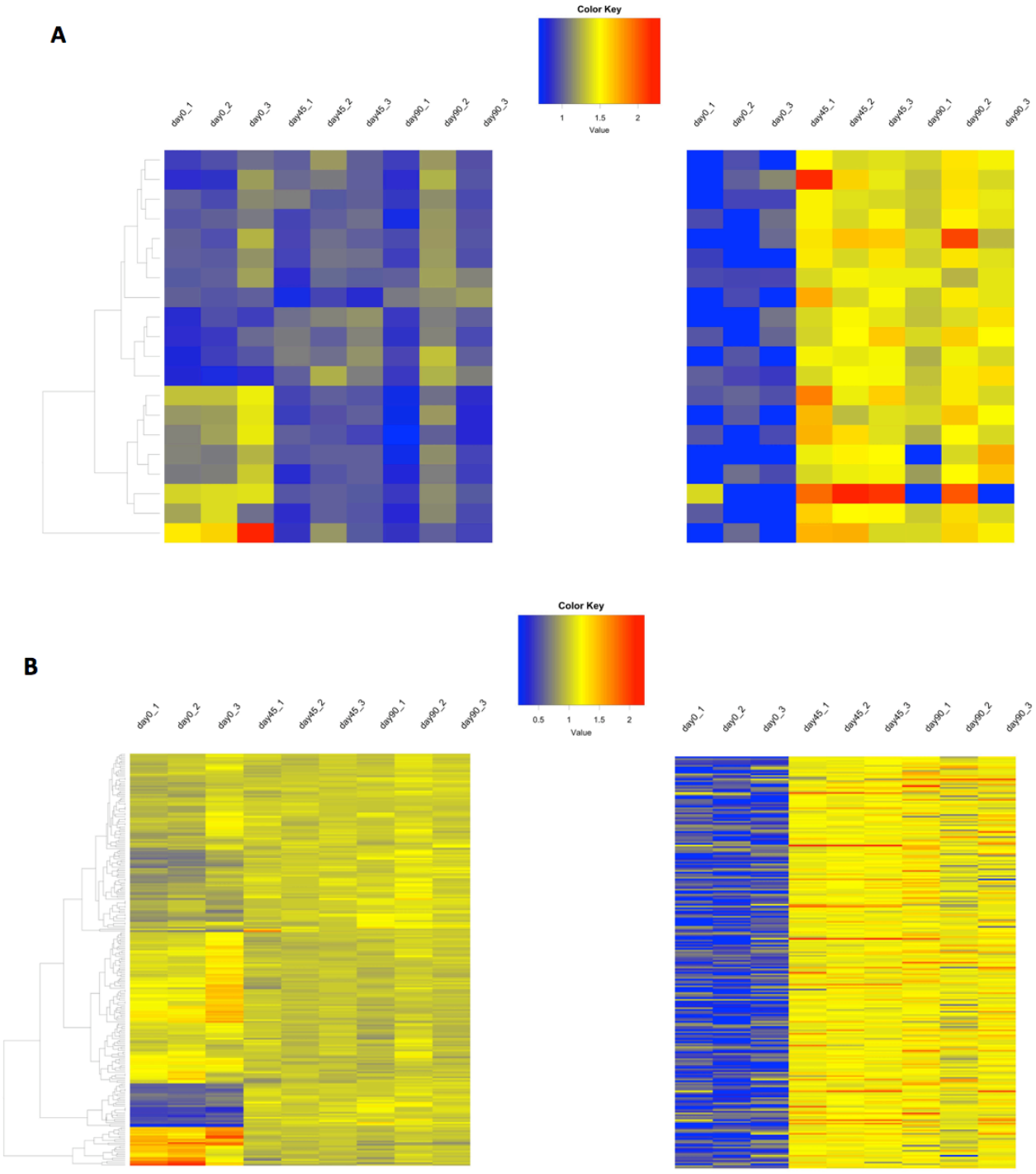


Figure 6.13. **Heat map of differentially expressed PTES transcripts**. Heat map showing expression estimates across differentiation of all transcripts reaching significance after enrichment analyses comparing A) day0 with differentiated samples (control) and B) day0 with IGF-1 treated samples. Canonical expression of genes with DE PTES (left) and PTES expressions (right) are shown for each comparison. Each row shows PTES expression across time points and total canonical junction expression from PTES host gene. Blue colour scale - from light to dark - indicates low expression to high expression, for each transcript tested.

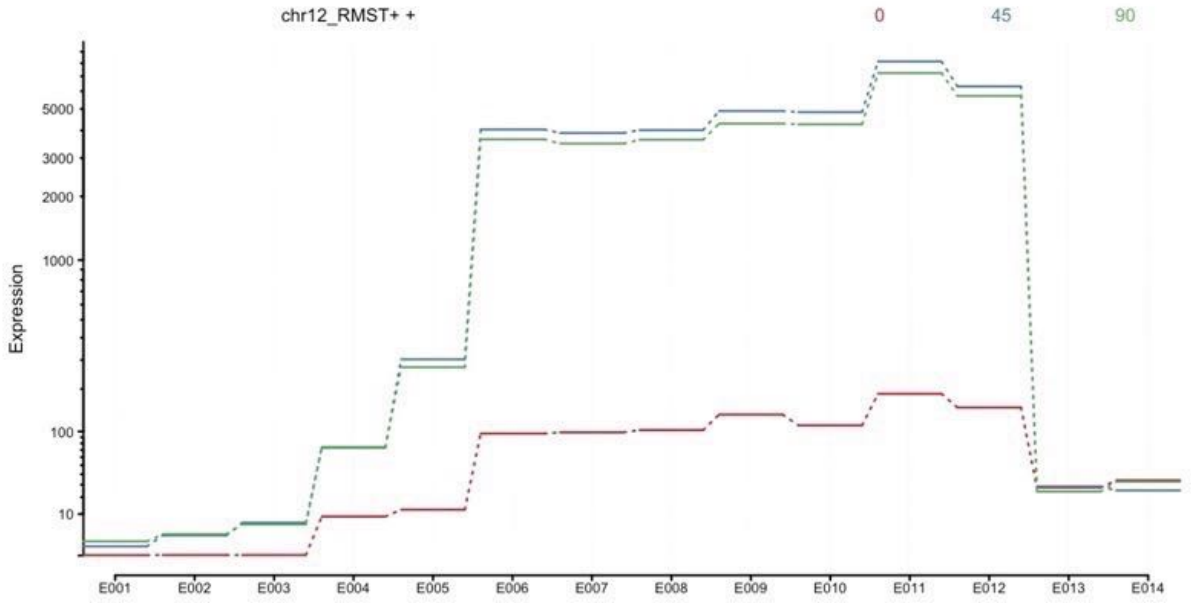
Typically, enrichment analyses are followed by gene ontology analyses to infer functional relevance of enriched transcripts. This, however, is of limited value to inferring functional

significance of PTES transcripts. The handful of PTES transcripts with known functions, play roles different from their cognate linear transcripts (Memczak et al., 2013; Hansen et al., 2013). Nevertheless, to identify biological pathways consisting of PTES genes with significantly enriched PTES transcripts, public gene ontology tools (Enrichr (Chen et al. 2013) & WikiPathways (Kutmon et al. 2016)) were interrogated. Expectedly, results (see appendix 9.5) show most of these genes (with PTES transcripts enriched in differentiated samples) to be involved in signaling pathways, and subsequent differentiation.

6.3.5 PTES transcript from RMST is circular and increases in abundance upon differentiation

The DE transcripts include 266 distinct PTES transcripts from 241 PTES producing genes, with 239 of these genes coding for proteins and assumed to be functional. However, 2 DE transcripts are produced from 2 non-coding loci: *GUSBP2*, an unspliced pseudogene with no known function and *RMST*, a long non-coding RNA with functional relevance in neurogenesis (Ng et al., 2012 & 2013). Both genes have multiple PTES products, 4 from *GUSBP2* and 14 from *RMST*. One PTES transcript from the *RMST* locus, produced by a backsplice between exons 12 and 6 (RMST.12.6), has been reported to be a product of homotypic trans-splicing, thus, linear (Wu et al., 2013).

In Wu et al., (2013), RMST.12.6 was found to be the dominant transcript from the *RMST* locus in ESC and its abundance was observed to decrease upon differentiation, suggesting a role in pluripotency maintenance. Figure 6.14 (inset table) shows total read counts observed for this transcript in each time point. Read counts are the lowest in day 0 samples, relative to other time points. This PTES is also found to be significantly enriched upon differentiation (p-value: 2.48×10^{-16}). On the evidence of read counts across exons in this gene (Fig. 6.14), the majority of reads emanate from exons predicted to be within a circular product (exons 6 - 12) and there are virtually no reads external to the circRNA, inconsistent with a trans-spliced PTES product. Alternative linear isoforms may contribute to this difference in read depth, nevertheless, we would expect ~2:1 ratio between PTES exons and non-PTES exons for linear PTES transcripts, where repeated exons contribute ~2X more reads. On the contrary, there is a ~10X excess of reads relative to terminal exons in day 0; and over ~600X more reads in higher time points. Furthermore, there is no RefSeq or GENCODE annotated linear isoform comprised only of exons 6 - exon 12, that may explain the observed read depth difference.



Samples	PTES Count	Canonical Junctions Count	Abuncance Ratio (P / C)
day 0	96	731	0.131
day 45C	5,410	23,206	0.233
day 45IGF	6,155	27,959	0.220
day 90C	4,331	19,941	0.217
day 90IGF	6,354	28,330	0.224

Figure 6.14. **Read distribution of RMST across time points.** Distribution of reads in RMST locus across differentiation series (days 0, 45 and 90 control samples). Inset) PTES and canonical junction read counts observed in respective samples.

To confirm these observations, real-time PCR of RMST.12.6 was performed in replicates at days 0 and 30 (Fig 6.15). The increase in abundance during differentiation is clear, with RMST.12.6 registering over 6 cycles earlier in day 30 samples than in day 0. To investigate the structure of RMST.12.6, RNA extracts from day 0 were treated with RNase R to selectively digest linear molecules. Real-time PCR experiments were then performed, comparing the effect of RNase R digestion on RMST.12.6 relative to a canonically spliced junction (RMST.12.13) and a housekeeping linear transcript, *PPIA*. Results in Fig. 6.16 shows noticeable variation in expression for the control assays, when treated and untreated samples were compared. However, this is not the case for RMST.12.6, as any variation is reduced, relative to that of control assays (inset, Figure 6.16). As abundance of RMST.12.6 is not altered upon RNase R digestion, the transcript is circular and cannot have arisen from trans-splicing as previously suggested (Wu et al., 2013).

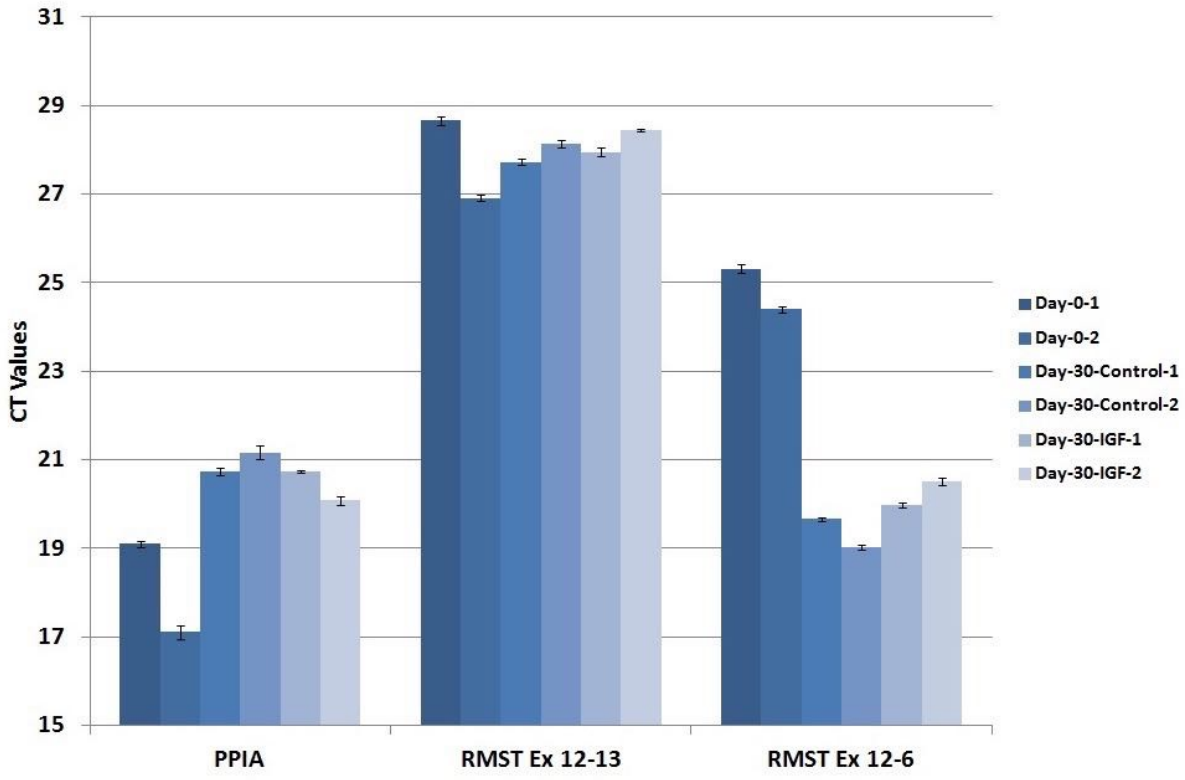


Figure 6.15. **Experimental confirmation of RMST.12.6**. Quantitative PCR (qPCR) results of RMST.12.6 at 2 time points - days 0 and 30. Lower CT values in day 30 indicates higher expression of RMST.12.6 at that time point, relative to that observed in day 0. There is little of no difference in expression between samples incubated with and without IGF1. Unpublished qPCR results obtained from Dr. Alhassan (Newcastle University, UK) presented here.

Although Wu et al., (2013) characterized this transcript from PolyA+ RNAseq data, their *in vitro* confirmation using RNase R digested sample failed to produce the expected amplicon size. This could be due to the linearisation of the circular product, as have been reported by other studies (Jeck et al., 2013). Notably, concentrations of RNase R used in Wu et al., (2013) were higher than commonly used in studies characterizing circRNAs (Memczak et al., 2013; Salzman et al., 2013), possibly resulting in sensitivity to exonuclease activity. To date, the only other ESC time course which includes total RNAseq data, mined for circRNAs comes from mouse (Rybak-Wolf et al., 2015). Consistent with my results, the authors both identified PTES transcripts corresponding to the human RMST.12.6 in mouse, found them to be enriched in RNase R digested mouse samples and reported higher counts in day 12 (123 reads) of a neuronal differentiation programme, compared to 0, 5 & 24 reads in days 0, 2 & 4 respectively (see Table S4, Rybak-Wolf et al., 2015).

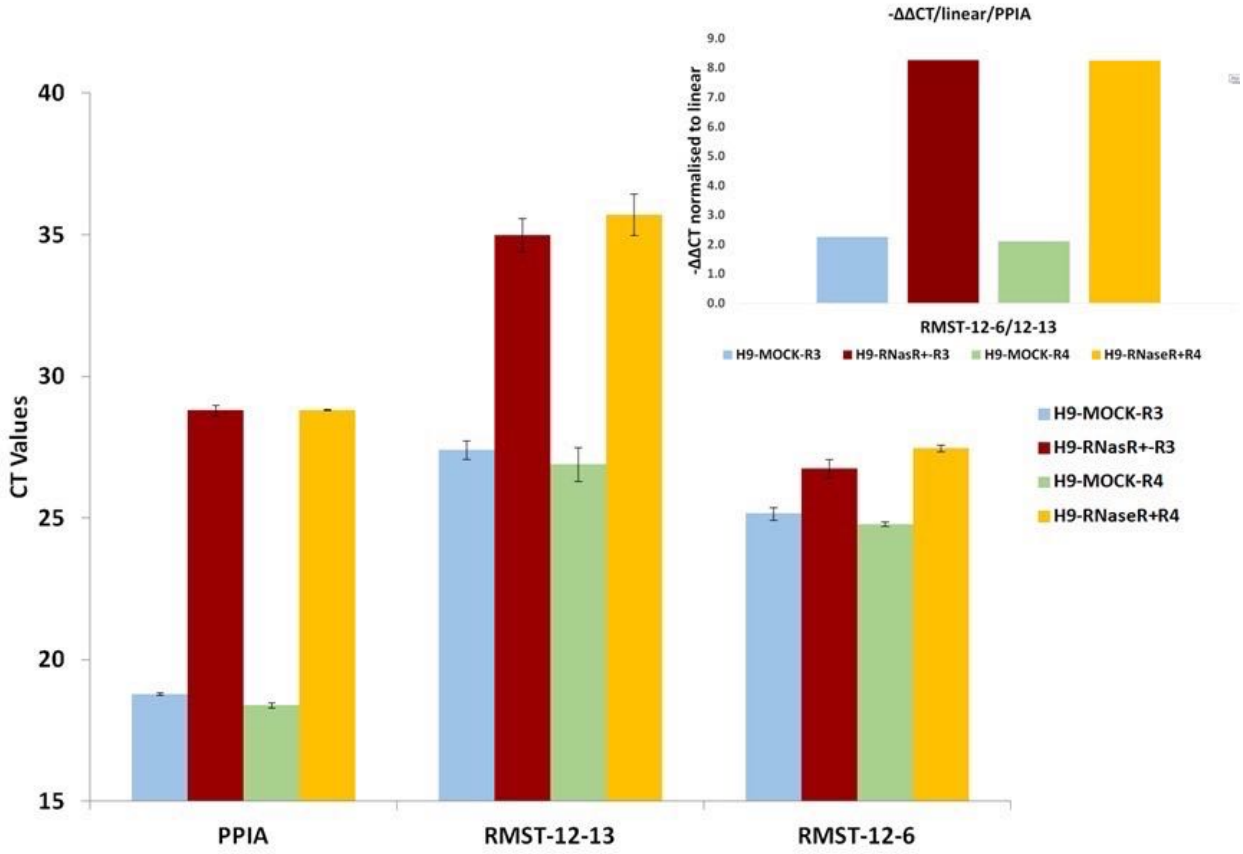


Figure 6.16. **Experimental confirmation of RMST.12.6 circularity**. qPCR results of PTES and canonical junction from *RMST* at day 0, with and without RNase R digestion. *PPIA*, a housekeeping gene is shown as control. Inset) Effect of RNase R treatment is shown by normalising CT values of RMST.12.6 to that of linear canonical junction (RMST.12.13) and *PPIA*. Unpublished qPCR results obtained from Dr. Alhassan (Newcastle University, UK) presented here.

Another result critical to conclusions made in Wu et al., (2013) include comparisons between expression of the PTES and the canonical transcript. Figure 6.17 below shows the positions of primer sequences used in that study. It is apparent that the primer pairs designed to amplify the linear transcript are internal to the predicted circular product, thus, their expression estimates are unlikely to be independent. This raises further questions about the reported switch in isoforms (PTES-to-linear) upon differentiation, as both primer sets amplify the RMST.12.6 circRNA. As my qPCR results show, RMST.12.6 is indeed more expressed in all time points, relative to the linear transcript, but the reported switch is not observed at 30 days or within 90 days of differentiation, a much longer time interval than the 21 days of differentiation in Wu et al., (2013). Taken together, our data indicates that RMST.12.6 is circular, and that it increases in abundance upon differentiation.

These results further suggest that RMST.12.6 is unlikely to contribute to pluripotency maintenance as previously reported (Wu et al., 2013). In contrast with functions defined for

RMST.12.6 in Wu et al. (2013), Ng et al., (2012 & 2013) found that transcripts from *RMST* did not associate with PRC2, interacted with *SOX2*, a transcription factor and regulated the expression of ~1000 genes via its association with *SOX2*. Knockdown of *RMST* resulted in less *SOX2* occupancy of many genes involved in neurogenesis, suggesting the role of *RMST* as a transcriptional co-regulator of *SOX2* and a possible guide RNA. Notably, manual analysis of siRNA sequence used in functional studies of *RMST* (Ng et al. 2013) revealed 100% identity to exon 9 (Fig. 6.17), within the predicted circRNA. As RMST.12.6 is the dominant transcript and increases upon differentiation, the reported functions of *RMST* in neurogenesis (Ng et al., 2013) may conceivably be ascribed to RMST.12.6, suggesting a previously unknown function for circRNAs.

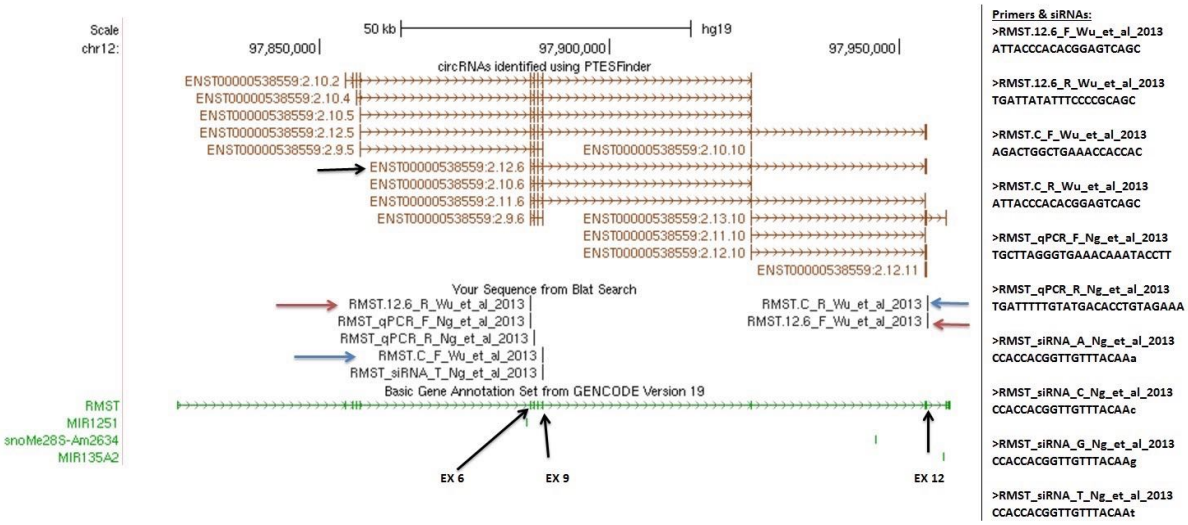


Figure 6.17. **Published RMST primers and siRNAs**. Primers used in Wu et al., (2013) to amplify PTES (red) and linear transcripts (blue). The expected amplicon from the linear transcript is within the predicted circRNA. SiRNAs used in functional analysis of transcripts from RMST locus and reported in Ng et al., (2013) are also shown. SiRNAs target an exon predicted to be included in RMST.12.6 circRNA. Multiple PTES produced from RMST locus are also shown (brown), and span from exons 2 to 12.

6.3.6 Multiple PTES transcripts originate from FIRRE and likely have previously unreported functional significance

The enrichment analyses performed above were designed to identify PTES transcripts enriched in specific time points during differentiation, independent of global and locus-specific changes of linear functional transcripts. While this design is appropriate for PTES transcripts originating from protein-coding genes, it may not necessarily be ideal for PTES from non-coding loci, which will not (by definition) be associated with functional coding transcripts and could lack any functional linear molecules. I therefore, re-examined the list of DE PTES transcripts from the first statistical analyses (sample and locus-level). Of the 3836 PTES

producing genes with transcripts reaching significance, 123 are non-coding and collectively produce 922 PTES transcripts. Ranking non-coding PTES producing genes by number of identified PTES per locus and total number of reads across time points, revealed that multiple transcripts originate from 110 genes (see appendix 9.5). Within this list are 12 long intergenic RNAs (Table 6.3), including *FIRRE* with 2857 PTES supporting reads from 20 distinct PTES transcripts (Fig 6.18A). PTES involving exons 10 and 5 (FIRRE.10.5) is observed with the highest number of supporting reads across all time points. Interestingly, expression estimates of FIRRE.10.5 are highest in undifferentiated samples, relative to differentiated samples.

Chromosome	Start	Stop	Gene	Gencode	Strand	PTES Reads	PTES Transcripts
chrX	130836677	130964671	FIRRE	ENST00000427391.1	-	2857	20
chr8	90729655	90769939	RP11-37B2.1	ENST00000504145.1	-	383	2
chr7	39773230	39832691	LINC00265	ENST00000340510.4	+	154	8
chr7	35791465	35840216	AC007551.3	ENST00000437235.3	-	140	2
chr1	149239867	149265510	RP11-403I13.4	ENST00000325963.8	+	98	1
chrX	102024106	102140334	LINC00630	ENST00000440496.1	+	67	5
chr4	119512927	119554884	RP11-384K6.6	ENST00000567913.2	+	66	5
chr3	197880120	197925886	FAM157A	ENST00000437428.2	+	43	4
chr21	29811666	30047170	AF131217.1	ENST00000433310.2	-	37	1
chr3	67705181	67998137	RP11-81N13.1	ENST00000482677.1	+	21	3
chr12	8448581	8549399	LINC00937	ENST00000544461.1	-	18	2
chr2	47441087	47572105	AC073283.4	ENST00000419035.1	-	11	1

Table 6.3. **List of non-protein coding genes with differentially expressed PTES.** Multiple PTES transcripts were identified from non-coding loci, including *FIRRE*, a long intergenic ncRNA known to have functional roles in regulating adipogenesis and maintaining DNA methylation in mouse fibroblasts.

To experimentally confirm the structure of FIRRE.10.5, Dr. Alhassan (IGM, Newcastle University) performed qPCR using RNA extracts from H9 ESC, treated with RNase R, an exonuclease that degrades linear molecules. Fig 6.15B shows that FIRRE.10.5 is resistant to exonuclease digestion and is circular.

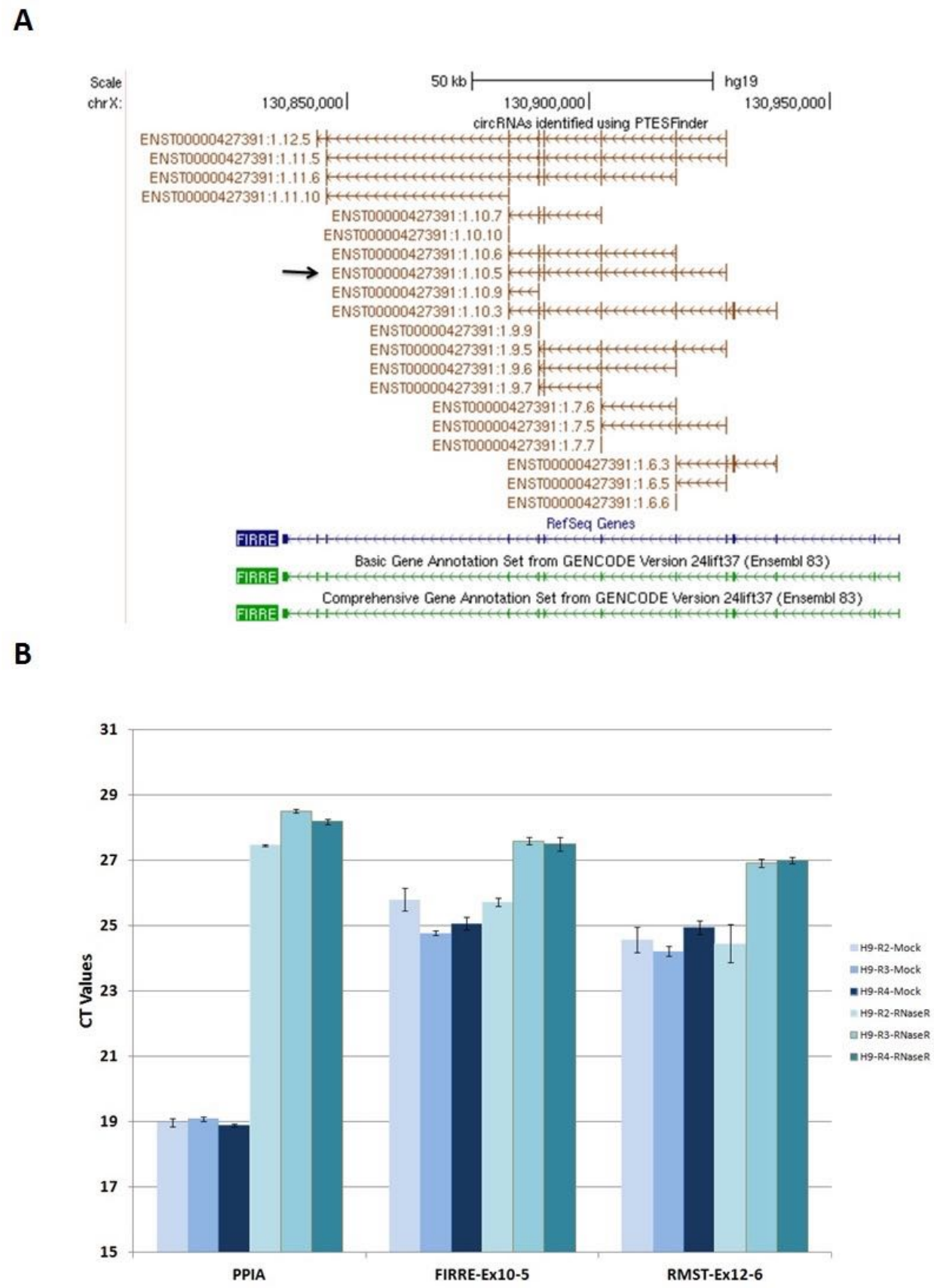


Figure 6.18. ***In silico* and *in vitro* confirmation of PTES from FIRRE.** A) Multiple PTES transcripts (brown) identified from human *FIRRE* locus, spanning from exon 3 to exon 12. B) Real-time PCR results of FIRRE.10.5 in H9 ESC, with and without RNase R digestion to confirm circularity. Unpublished qPCR results obtained from Dr. Alhassan (Newcastle University, UK) presented here.

Previous functional studies have shown transcripts from *FIRRE* to play roles in maintaining histone methylation in inactivated X chromosome (Yang et al., 2015) and in regulating adipogenesis (Hacisuleyman et al. 2014; Hacisuleyman et al. 2016). Local repeats within internal exons of *FIRRE* have also been shown to be highly conserved in mouse, and may aid localisation and interaction of multiple chromosomes (Hacisuleyman et al., 2014). As multiple PTES transcripts detected from this locus include exons containing these local repeats, I reasoned that multiple PTES from the same locus could conceivably be detected in mouse. To that end, I identified and screened 2 mouse embryonic stem cell datasets (Accessions: GSE47948 & GSE22959) for PTES. Six distinct PTES junctions: 20-4, 18-17, 17-17, 17-16, 15-6 and 12-12, were subsequently identified. Exons predicted to be within these circRNAs include known local repeats in this locus (Hacisuleyman et al., 2014). In characterising the functional role of transcripts from this locus, Yang et al., (2015) performed siRNA knockdown. BLAT analysis of siRNA nucleotide sequence used in that study revealed that exons predicted within identified PTES transcripts were targeted (Fig 6.19), raising the possibility that observed role in maintaining DNA methylation in inactivated X chromosome of mouse fibroblasts may be attributed to PTES from this gene. Additionally, the reported role in adipogenesis was observed in mouse ESC lacking all transcripts from *Firre*, including circRNAs (Hacisuleyman et al., 2014).

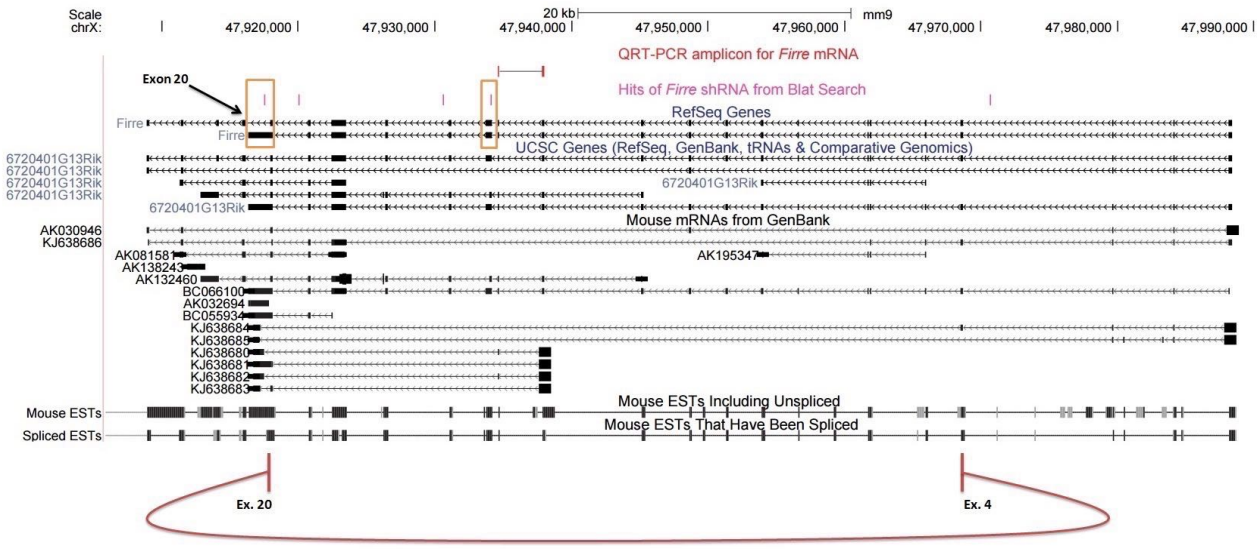


Figure 6.19. **Published siRNAs targeting transcripts from Firre gene in mouse**. SiRNAs used in Yang et al., (2015) target exons predicted to be included in circRNAs identified from *Firre*; including the terminal exon of the short RefSeq isoform (exon 20), which is alternatively spliced in other isoforms. The longest circRNA identified from mouse ESC samples is depicted (red). Figure adapted from Yang et al., 2015.

6.4 Discussion

PTES transcripts are non-coding RNAs, abundant in various human cell lines and tissues (Salzman et al., 2013); enriched in neurogenesis (Rybak-Wolf et al. 2015; Zhang et al. 2016; You et al. 2015) and epithelial-mesanchymal transition (Conn et al., 2015); and some exhibit marked increase in abundance during bone formation (Dou et al. 2016). Despite these reports, these transcripts remain without uniformly ascribed functional significance. Pluripotent cells are self-renewing and capable of being differentiated into any somatic cell (Morrison & Kimble, 2006; Boland et al., 2014). A core set of transcription factors are known to maintain the pluripotent state, by inducing epigenetic changes that suppress cell lineage commitment (Morrison & Kimble, 2006). Therefore, identifying PTES transcripts specifically enriched in ESC, may hint at potential roles in pluripotency maintenance, perhaps as miRNA sponges, acting to modulate expression of ESC-specific genes.

To explore their potential roles in pluripotency maintenance and cellular differentiation, H9 embryonic stem cells (ESC) were differentiated into retinal cells over a 90-day period, with and without incubating with IGF-1, a known facilitator of cellular differentiation (Mellough et al., 2015; Huat et al., 2014). To address the limitations of a previous attempt to identify ESC-specific PTES (Wu et al., 2013), RNA extracts of 3 biological replicates from each time point and treatment groups were sequenced. Screening for PTES in ribosome-depleted RNAseq samples from differentiation series (days 0, 45 & 90), the lowest number of transcripts were identified from day 0, representing ~3X less transcripts and supporting reads, relative to other samples. As three RNA binding proteins (RBP) (*QKI* [Conn et al., 2015], *MBNL* [Ashwal-Fluss et al., 2014] and *ADAR1* [Ivanov et al., 2015]) have been shown to regulate PTES biogenesis, their expression profiles were examined. Although expression of *MBNL* and *QKI*, known facilitators of PTES formation remained relatively stable upon differentiation, expression of *ADAR* was elevated in ESC. The multifunctional *ADAR* (Ota et al., 2013) has been shown to prevent the formation of secondary structures favourable to PTES biogenesis, by A-to-I editing that weakens double stranded RNA (Ivanov et al., 2015). Transcripts identified in differentiated samples, were found to be enriched for RNA editing sites within introns flanking backsplice junctions. Elevated expression of *ADAR* in ESC, likely contributes to the suppression of some PTES transcripts in pluripotent state, suggesting that most PTES may not be critical to self-renewal. CircRNA decay pathways may conceivably include endoribonuclease activity (Lasda & Parker, 2014). Profiling the expression of endoribonucleases, revealed slight elevation of transcripts from 2 genes (*DIS3* & *ZC3H12A*) in ESC, suggesting that endoribonuclease activity may contribute to lower PTES abundance observed.

Nucleosome position and intragenic CpG methylation have been linked to alternative splicing and transcriptional diversity during cellular differentiation (Maunakea et al, 2013, Gelfman et al., 2013, Singer et al., 2015). In my data, more canonical junction reads were observed in day 0, suggesting that many PTES producing genes are transcribed and non-detection of PTES from such genes in ESC may not be due to transcriptional silencing. Transcripts identified in this time point are characterized by shorter genomic distance between backsplice junctions and lower number of constituting exons. These observations are suggestive of slower transcription elongation rates of PTES genes in ESC. Bisulfite data from differentiation series of H9 ESC into retinal pigment epithelium (Liu et al., 2014), showed progressive reduction in levels of CpG methylation in exons. It is plausible that, as elongation rates increase due to intragenic epigenetic changes, distal splices sites and introns containing inverted repeats become available before sequential forward splicing completes (depicted in Fig 6.20). Such phenomenon may explain the reported correlation between alternative splicing and PTES (Surono et al., 1999; Kelly et al., 2015).

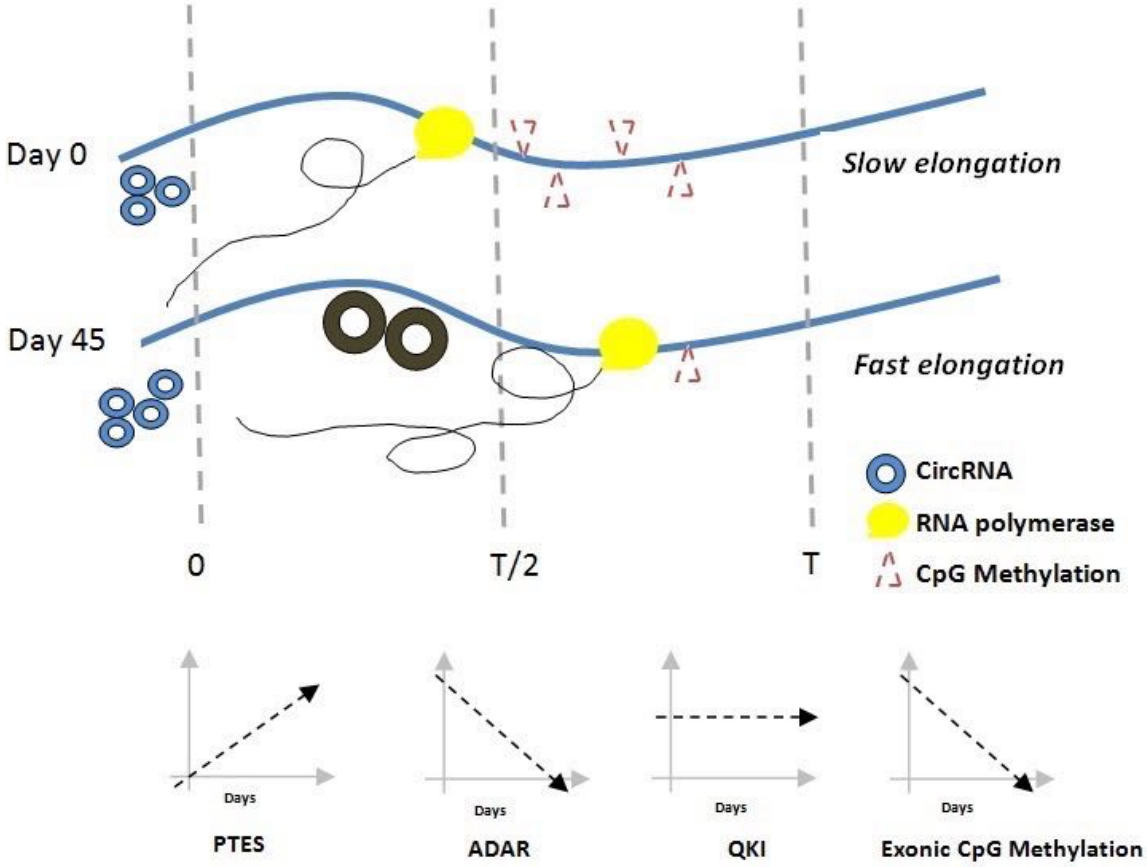


Figure 6.20. **Long range intron pairing is facilitated by transcription elongation rate**. Difference in unspliced size and number of exons predicted within circRNAs upon differentiation is likely due to changes in CpG methylation and subsequent increase in transcription elongation rate. In this model, time for transcription to complete (T) is reduced upon differentiation, as CpG methylation reduces, resulting in additional PTES events. Generally, increase in PTES abundance correlates with decrease in expression of *ADAR*, an RNA editing enzyme and intragenic CpG demethylation.

Profiling expression changes of PTES and canonical junctions between time points, a striking correlation was observed. My results showed that PTES levels increase 2X faster than canonical junction levels in differentiated samples. This higher rate of change in PTES expression is presumably as a result of differential stability and subsequent accumulation (Westholm et al., 2014; Rybak-Wolf et al., 2015; Enuka et al., 2015). This, however, suggests that PTES transcripts may be less regulated than cognate mRNAs and are indeed by-products of changes in transcription.

Enrichment analysis identified no PTES transcripts enriched in ESC, and no apparent short-term effect of IGF-1 treatment on PTES abundance, consistent with reports in Enuka et al., (2015). Taken together with their suppression in ESC, presumably due to combinatory effects of 1) elevated *ADAR* expression 2) elevated expression of endoribonucleases and 3) progressive epigenetic changes upon differentiation; results strongly suggest that PTES transcripts are inconsequential to pluripotency maintenance. These results are at odds with an earlier report of PTES from *RMST* locus with expression pattern suggestive of role in pluripotency (Wu et al., 2013). Contrary to that report, my results confirm this transcript to be circular and expectedly increasing in expression upon differentiation. Functional analysis of transcripts from *RMST* by Ng et al., (2013) revealed important roles in neuronal differentiation and interaction with *SOX2*, a transcription factor. Manual examination of siRNA used in that study however, revealed exons predicted within RMST.12.6 as targets of the knockdown, suggesting that PTES from this locus may have functional relevance in neuronal differentiation, guiding *SOX2* to promoters of genes necessary for brain development.

Finally, multiple unreported PTES transcripts from *FIRRE*, another lncRNA, were identified (in both hESC and mESC), some differentially expressed upon differentiation. Previous knockdowns of transcripts from *FIRRE* revealed roles in maintaining histone methylation (Yang et al., 2015) and regulating adipogenesis in mouse ESC by unknown mechanisms (Hacisuleyman et al., 2014). SiRNAs used in Yang et al., (2015) target exons predicted within circRNAs from *Firre* and suggest that circRNAs from this locus may be functional. Taken together with potential functional relevance of RMST.12.6, circRNAs from non-coding genes may have functional significance in differentiation, different from established functions of 3 circRNAs as miRNA sponges.

6.5 Conclusion

In this chapter, I assessed the distribution of PTES transcripts in differentiation series of H9 ESC, to identify transcripts with potential roles in pluripotency maintenance and differentiation.

PTES biogenesis and abundance were found to be lowest in ESC, and I presented evidence suggesting that suppression by RNA editing, endoribonuclease activity and CpG methylation may contribute to this. Following differential expression analyses, no PTES transcript was found to be significantly enriched in ESC, relative to differentiated samples, suggesting PTES from protein-coding genes to be non-essential to pluripotency maintenance. PTES transcripts, however, accumulate during differentiation, tracking canonical expression changes and as a mechanism, may compete with alternative splicing. Contrary to a previous report (Wu et al., 2013), my results showed that PTES from *RMST* (Ex 12-6) is circular, increases in expression upon differentiation and potentially functions as a co-transcriptional regulator via its association with *SOX2*. PTES transcripts from another non-coding gene, *FIRRE*, were found to decrease upon differentiation, circular and may be functional.

Chapter 7. General discussion and future work

7.1 General discussion

Many eukaryotic genes undergo splicing to join exons sequentially, in an order consistent with their order in the genome. Although exons can be alternatively spliced, the exon arrangement in mature transcripts is typically consistent with that in the genome, with upstream exons spliced to downstream exons. The high efficiency of splicing, coupled with mechanisms for degrading mis-spliced transcripts minimise potential deleterious effects of defective transcripts. The complexity of eukaryotic transcriptomes is however enhanced by the existence of re-arranged transcripts and other chimeric transcripts, without known impact on cellular activity. Post-transcriptional exon shuffling (PTES) has now been shown to result in both linear and circular RNA molecules, and progress has been made in characterising PTES events in various cell lines and tissues. Despite this progress, when I started my research, little was known about their mechanisms of formation, how they are regulated, nuclei-cytoplasmic export mechanisms, spatio-temporal distributions or their functional significance. This thesis focussed on elucidating some of the mysteries surrounding these transcripts, using *in silico* and *in vitro* approaches.

Transcripts arising from PTES eluded detection using *in silico* and *in vitro* methods that were biased towards identifying established or well characterized RNA species, specifically mRNAs. Recently described computational methods for PTES detection typically reduce the likelihood of false positive identifications by first aligning all reads to the genome and restricting detection to reads unaligned to the genome. The rationale for this approach is that: reads aligned to other genomic features cannot be used to correctly characterise PTES events and reads not aligned to the genome will be enriched with reads in inverted order relative to the genome. This approach however, overlooks other sources of artefacts, including reads emanating from tandem-exon repeats, unmapped segmental duplications, read through transcripts and template-switching artefacts.

Critical to research of these novel transcripts is an accurate identification method that reduces misidentification of artefacts as PTES. My method, PTESFinder, is equipped with filters that directly target these sources of artefacts. In the first results chapter, the efficacy of these filters was assessed, revealing that all filters excluded an overlapping population of reads, likely to be filtered by other published methods. However, distinct populations of false positive

reads are excluded by each filter, highlighting their usefulness. Comparisons with other published tools using both simulated and real RNAseq data revealed that PTESFinder has the highest specificity and comparable sensitivity.

One limitation of PTESFinder is its reliance on curated linear transcript annotations with known splice junctions. This limitation reduces the number of identifiable PTES to only re-arrangements at known exonic regions, overlooking PTES from intergenic or intronic regions. The reliance on known splice sites however, contributes to the high specificity of PTESFinder. It is now known that the vast majority of PTES occur at known splice junctions (Jeck et al., 2013; Liang & Wilusz, 2014), thus limiting the effect on sensitivity of reliance on annotated transcripts. Further analysis using annotation-free versions of PTESFinder, revealed that most identifications from published annotation-free methods are replete with false positives, particularly, methods relying on non-splice aware aligners. This observation conceivably contributes to the reported discordance in PTES identified from the same samples using different methods (Yu et al., 2014; Hansen et al., 2015).

An unanswered question pertaining to PTES is that of their export to the cytoplasm from the nucleus. Studies have reported the identifications of PTES in the cytosol (Salzman et al., 2012), and it has been suggested that they exit the nucleus during cell division (Jeck et al., 2013). Prior to my study, no transcritptome-wide quantitative analysis of PTES populations in sub-cellular compartments had been reported. Thus, it remained unclear if a nuclei-cytoplasmic pathway existed for PTES transcripts and why they were actively exported to the cytosol if they were indeed products of defective splicing. My analysis of RNAseq from various cellular compartments of 7 human cell lines revealed PTES distribution patterns inconsistent with efflux to the cytosol during mitosis. I identified a variety of PTES events in the nuclei, some of which are retained in the nucleus, with no detectable reads supporting their presence in the cytosol of all cell lines. More interesting was the identification of incompletely processed PTES transcripts with retained introns abundant in the nucleus. One example of a PTES with both intron-containing and intronless circular isoforms is CAMSAP1.3.2, previously reported by Salzman et al (2013) and Zhang et al., (2014). My results showed that the intron-containing isoform is released from chromatin, translocated from the nucleoplasm to the nucleolus, suggesting that processing of this PTES is incomplete during transcription and may continue in another cellular compartment. This conclusion is supported by findings in Zhang et al., (2014), showing that this isoform is unstable with a half-life of 7 mins, presumably as a result of further splicing or rapid linearisation and subsequent decay.

The question about whether PTES formation is indeed post-transcriptional or occurs during transcription is a subject of intense debate. *In vitro* experiments using mini-gene constructs have

been performed by others, showing PTES to either require the formation of polyA tails on nascent transcript (Liang & Wilusz, 2014) or occur co-transcriptionally (Ashwal-Fluss et al., 2014; Kramer et al., 2015); the latter is consistent with most splicing (presumably including back-splicing) occurring during transcription (Ameur et al., 2011; Tilgner et al., 2012; Girard et al., 2012). It is however known that depending on the rate of transcription elongation and size of transcript, splicing may be committed but not completed during transcription (Bentley 2014).

My analysis of PTES in chromatin-associated RNAs identified ~2000 PTES transcripts, providing initial evidence of PTES formation during transcription. Further *in silico* analysis of co-transcriptional splicing by comparing reads from pre-mRNAs to that of mature mRNAs, revealed that, as a group, exons from PTES producing genes likely undergo more co-transcriptional splicing than other genes. Interestingly, the rate of intron removal during transcription was observed to reduce significantly in PTES producing genes relative to other genes, when the first and last intron removal rates were compared. Reasons for this observation are not readily available without *in vitro* experiments. Nevertheless, it is conceivable that for some transcripts, following PTES during transcription and subsequent release from chromatin, spliceosomal assembly on remnants of the nascent transcript is impeded, resulting in retention at nuclear speckles. This may explain the rapid removal of upstream introns but the observed lower rate of last intron removal. As exemplified by CAMSAP1.3.2, no reads supporting the first intron were detected; the peak read density observed across this gene was observed between exon 2, intron 2 and exon 3, underscoring the contribution of PTES to reads from this gene. Release of CAMSAP1.3.2 from chromatin is then followed by reduced rate of last intron removal, suggesting its retention. My findings are consistent with reports of competition between PTES and forward splicing (Ashwal-Fluss et al., 2014); and imply that PTES can impact the expression of cognate linear transcripts.

Enrichment analysis of PTES populations in the nucleus and the cytoplasm further revealed that ~9% of PTES are enriched in the cytosol relative to the nucleus. This enrichment in the cytosol is likely due to differential stability of circRNAs, which are resistant to exonuclease activity. This however, raises the question of their relevance in the cytosol. Assessing the abundance of PTES in RNAseq data from sucrose gradient fractions of HEK293 with and with arsenite treatment to inhibit translation, revealed that these transcripts are typically not bound by ribosomes and do not contribute to the proteome. However, there may exist, some linear PTES transcripts not subject to nonsense mediated decay that can plausibly produce proteins. Nevertheless, my results showed that, consistent with other reports (Jeck et al., 2013; Guo et al. 2014), PTES do not contribute significantly to the proteome.

Like other RNA species, various factors including rate of transcription, export from the nucleus and degradation are likely to affect the abundance of PTES transcripts. Having observed interesting patterns of spatial PTES distribution in sub-cellular compartments, I extended my investigation to anucleate cells to assess PTES distributions in a system not affected by steady-rate transcription and export levels. Translation (Weyrich et al. 2009), cytoplasmic mRNA splicing (Denis et al., 2005) and miRNA biogenesis (Landry et al., 2009) have been reported in platelets, further raising the possibility of identifying novel PTES events within these anucleate cells.

In platelets, I identified many PTES junctions, including a large number that had not been identified (or reported) in nucleated cells. Comparisons of PTES abundance in platelets with that of PTES in nucleated tissues revealed that platelets are enriched for PTES, 17 to 188-fold compared to nucleated samples and 14 to 26-fold relative to nucleated samples treated with RNase R to selectively remove linear RNAs. The proportion of reads supporting PTES in platelets were found to be ~240X more than observed for sub-cellular compartments of nucleated human cell lines. These striking observations raised questions about circRNA biogenesis in platelets. However, with the exception of PTES from platelets-specific genes, the vast majority of PTES identified in platelets have previously been reported. Although splicoesomal proteins are known to be present in platelets (Denis et al., 2005), reads from genes reported to undergo splicing in platelets were not easily detectable. Similarly, I established that the high abundance of PTES seen in platelets is observed in other anucleate cells, as I identified many PTES transcripts from mature erythrocytes, further dispelling the notion that PTES enrichment in platelets may be due to cytoplasmic splicing events.

As platelets are short lived, and their lifespan will be influenced by the availability of full length mRNA transcripts inherited from their megakaryocyte progenitors, I investigated the possibility that the observed PTES enrichment is due to RNA decay of linear molecules and stability of circRNAs. Comparing expression across exons predicted within circRNAs to exons external to circRNAs, I found that in platelets, circRNA exons are significantly more enriched than in nucleated tissues. For some genes, the contribution of exons within circRNAs is 99%, thus, deplete of reads in exons external to circRNAs. *Ex vivo* and *in vitro* analysis of RNA decay confirmed the same pattern of circRNA enrichment relative to cognate linear transcripts.

Strikingly, the effect of RNA decay on circRNA enrichment in platelets is significantly higher than is obtainable for samples treated with RNase R to enrich circRNAs. This observation raises questions about RNase R treatment as a source of technical variation in estimating the abundance of PTES. The extent of linear RNA degradation observed in platelets also highlighted the possibility of circRNA contamination in polyA+ RNA fractions. PTES

identified from polyA+ platelet samples are characterized by high composition of adenosine residues, suggesting their pull-down during mRNA isolation. It is currently not clear how circRNA contamination impacts expression estimates of cognate mRNAs in differential expression studies.

Three studies have demonstrated that some circRNAs can compete with linear RNAs and act as decoys for miRNAs (Hansen et al., 2011 & 2013; Memczak et al., 2013; Zheng et al., 2016). However, in one report miR-671 was found to induce the degradation of circRNA from *CDR1*, suggesting a degradation pathway for PTES (Hansen et al., 2011). A striking difference in number of PTES identified from 2 male platelets samples was observed and was not readily explainable by sampling, technical variation or the limited demographics information of blood donors available.

I investigated this fluctuation in identified PTES further, by assessing read distribution patterns across inferred circRNA sequence and found patterns suggestive of miRNA-induced degradation. For some circRNAs, sequence regions were deplete of read coverage and some of these regions correspond to known miRNA binding sites. By implication, circRNAs are available substrates for endoribonucelases (facilitated by miRNAs) upon depletion of linear molecules. As circRNAs are less abundant than mRNAs, this premise portrays mRNAs as decoys for miRNA binding. It follows that for circRNAs, any miRNA sponging effect is minimal, considering their relative low expression and their effectiveness as sponges will be miRNA-dependent. In the case of circCDR1 for instance, its effectiveness as a miRNA-sponge will likely depend on the concentrations of miR-7 and miR-671, with both competing for binding sites within circCDR1 and the binding of one (miR-671) inducing degradation. Notably, studies showing circRNAs as potential miRNA sponges have characterized this effect based on exogenous levels of circRNA expression (Hansen et al., 2011; Memczak et al., 2013; Zheng et al., 2016), which are likely to be inflated when compared to endogenous expression levels.

In my final results chapter, I investigated the temporal distribution of PTES upon cellular differentiation of H9 embryonic stem cells (ESC). I aimed to identify factors that may influence changes in expression and abundance of PTES during differentiation, and infer functional relevance of PTES based on differential expression across time points. Previously, expression changes between developmental stages have been used to infer functional significance of specific PTES (Wu et al., 2013), but without controlling for both global and locus-specific changes in transcription changes. My results showed that the expression changes of PTES track those of linear canonical transcripts, but that the rate of increase in expression levels is higher for PTES than for linear transcripts. This variation in rates is likely due to accumulation over

time, as a result of their higher stability relative to linear transcripts. More PTES were identified from differentiated samples than from ESC, presumably due to changes in CpG methylation across exons and expression of various RNA binding proteins (RBPs) and endoribonucleases. Two RBPs have been shown to facilitate PTES (*QKI* [Conn et al., 2015], *MBNL* [Ashwal-Fluss et al., 2014]), but my results showed that their expression patterns are relatively stable across time points, suggesting that they may not be critical to suppression of PTES in ESC. However, expression levels of *ADAR1* and 2 endoribonucleases (*DIS3* & *ZC3H12A*) were found to be highest in ESC, decreasing upon differentiation, thus, inversely proportional to the number of PTES identified across time points. *ADAR1* has been shown to inhibit the formation of double stranded RNA and subsequent PTES formation. Flanking intronic sequences of PTES junctions identified in differentiated samples and not in ESC were found to contain significantly higher numbers of RNA editing sites, relative to PTES found in ESC. This is consistent with *ADAR1* inhibiting intron pairing by A-to-I editing (Ivanov et al., 2015), thus, suppressing PTES in ESC.

Characteristics of PTES identified from both stages (differentiated and undifferentiated) were observed to differ. PTES identified in differentiated samples were shown to include more exons in inferred circular transcripts and their unspliced sizes higher than observed for PTES in undifferentiated samples. Coupled with an inverse relationship between intragenic CpG methylation and number of PTES identified across time points, these observations are suggestive of increased transcription elongation rates, making distal splice sites available for PTES. This premise is supported by a recent study that found that more PTES were identified with a variant of RNA polymerase 2 with faster elongation rates (Zhang et al., 2016). Although progress has been made in elucidating mechanisms of PTES formation, my results show another layer of regulation for PTES. Prior to this study (and that of Zhang et al., (2016)), it was unclear what signals affect the choice of whether a nascent transcript results in PTES of canonical transcript and how multiple PTES originate from same locus. Like alternative splicing, splice site competition and intron-pairing patterns are likely dependent on transcription elongation rates. This is likely to contribute in the reported tissue-specific PTES expression variations (Salzman et al., 2013), as various cells have different transcription elongations rates.

Enrichment analysis to identify differentially expressed PTES transcripts from protein-coding genes, collectively resulted in less than 300 PTES across all time points. Expectedly, these PTES increased in expression upon differentiation, consistent with accumulation. Contrary to a previous report (Wu et al., 2013), no PTES was found enriched in ESC, relative to other time points. Such enrichment would be taken as evidence of functional relevance in pluripotency maintenance. Furthermore, contrary to the report of Wu et al., (2013), a PTES from *RMST* locus (RMST.12.6) was found to be circular, the dominant transcript from the locus

in all time points and to increase in abundance upon differentiation, not decrease as previously reported. Previous functional analyses of transcripts from this locus established an association with *SOX2* (a transcription factor) and possible roles in neuronal differentiation (Ng et al., 2013). Manual examination of siRNAs used in knockdowns in these studies (Ng et al., 2013; Wu et al., 2013), showed that exons included in the circRNA were targeted and any functional significance previously ascribed to the linear transcript can conceivably be that of the circRNA.

Similar results were obtained for circRNAs from *FIRRE*, a long non-coding gene, with transcripts associated with maintaining DNA methylation in mouse fibroblasts (Yang et al., 2015) and possible roles in regulating adipogenesis in mouse ESC (Hacisuleyman et al., 2014). Moreover, knockdown of *Firre* in mouse ESC was reported to induce the down regulation of genes involved RNA processing (Bergmann et al. 2015). As siRNAs used in these functional studies target exons within circRNAs from *Firre*, circRNAs may have previously unreported functional significance.

7.2 Conclusions and Future Work

Various studies have identified features that promote PTES formation using minigene constructs, comprising only of backspliced exons and flanking introns (Ashwal-Fluss et al., 2014; Liang and Wilusz 2014; Kramer et al., 2015; Starke et al., 2015). As this approach introduces exogenous RNAs by transfection and full length genes are not used, PTES events observed are not produced in the accurate context of the transcriptome under study. This approach may not reflect the competition for spliceosomal proteins and overlooks the effect of epigenetic regulation on PTES formation. Recent advances in genome editing using the CRISPR/CAS9 system (Kim 2016) can be used to overcome some of these limitations. In this system, specific genomic features can be targeted by RNA-guided endonuclease cleavage and editing (Kim, 2016), allowing for *in vivo* investigation of PTES formation. Genome editing affecting PTES formation may induce phenotypic changes and negate post-transcriptional siRNA knockdowns for investigations of functional relevance.

Second, the identification of a nucleo-cytoplsmic export pathway for PTES is needed and may shed light on their relevance. It is currently not clear why some PTES are retained in the nucleus. For intron-containing PTES, it is conceivable that they retain signals that confine them to the nucleus. However, some mono-exonic PTES transcripts may not retain those signals and reason for their abundance in the nucleus or lack of export to the cytosol is a mystery. *In vitro* experiments perturbing well characterized export pathways may provide insight on how such perturbations affect PTES.

Third, the availability of computational methods for PTES identification from high throughput RNAseq data offers opportunities for further studies to characterise these transcripts. However, the lack of adopted standards for detection and annotation of these transcripts presents challenges in comparing results from various studies. Reported discordance in results obtained from same sample using different methods (Yu et al., 2014), and the observed differences in specificities and sensitivities of various tools, strengthens the need for specific standards. Further studies of PTES may also benefit from a repository of PTES identified from various organisms, with a single computational tool and accepted annotation standards. Of note is a single repository - circbase.org (Glazer et al., 2014); this database currently catalogs published PTES, identified using various methods, thus, contain an unknown number of false positive predictions. I propose a single method or an ensemble of methods to be used to screen RNAseq data of a pre-defined standard, quantifying both PTES and canonical junction counts and generating other metrics associated with identified PTES. Similar to efforts by GENCODE and RefSeq, this will undoubtedly increase the confidence in PTES predictions and allow for sophisticated quantitative analyses of archived results.

The quantitative analysis of PTES and linear transcript in platelets raised additional questions about the composition of platelets transcriptomes. Previously reported pre-mRNA transcripts that apparently undergo cytoplasmic splicing (Denis et al., 2005) were not detected in data analyzed. The non-detection of these transcripts is consistent with the extensive degradation of linear RNA molecules observed and raises questions about the extent of cytoplasmic splicing in platelets. A comprehensive assessment of all RNA species, including capped and polysome-associated transcripts may further our understanding of the platelets transcriptomes.

Finally, studies have shown that degraded samples can yield inaccurate expression estimates and subsequent false positive differential expression predictions, prompting the need for *in silico* correction of expression estimates (Romero et al. 2014; Wang et al. 2016). The observed contamination of polyA+ fractions by circRNAs may likely impact expression estimates and downstream statistical analyses. It may be necessary to first assess the impact of this contamination of expression estimates and provide a framework for quantifying circRNA contamination and accounting for this effect on RNA expression estimates. However, this may only be necessary if samples to be compared vary significantly in quality and RNA integrity.

Chapter 8: References

Adiconis, Xian, Diego Borges-Rivera, Rahul Satija, David S DeLuca, Michele a Busby, Aaron M Berlin, Andrey Sivachenko, et al. 2013. "Comparative Analysis of RNA Sequencing Methods for Degraded or Low-Input Samples." *Nature Methods* 10 (7): 623–29. doi:10.1038/nmeth.2483.

Afgan, Enis, Dannon Baker, Marius van den Beek, Daniel Blankenberg, Dave Bouvier, Martin Čech, John Chilton, et al. 2016. "The Galaxy Platform for Accessible, Reproducible and Collaborative Biomedical Analyses: 2016 Update." *Nucleic Acids Research* 44 (W1): W3–10. doi:10.1093/nar/gkw343.

Akiva, Pinchas, Amir Toporik, Sarit Edelheit, Yifat Peretz, Alex Diber, Ronen Shemesh, Amit Novik, and Rotem Sorek. 2006. "Transcription-Mediated Gene Fusion in the Human Genome." *Genome Research* 16 (1): 30–36. doi:10.1101/gr.4137606.

Al-Balool, Haya H, David Weber, Yilei Liu, Mark Wade, Kamlesh Guleria, Pitsien Lang, Ping Pitsien Lang Ping Nam, et al. 2011. "Post-Transcriptional Exon Shuffling Events in Humans Can Be Evolutionarily Conserved and Abundant." *Genome Research* 21 (11). Cold Spring Harbor Laboratory Press: 1788–99. doi:doi: 10.1101/gr.116442.110.

Altschul, S F, W Gish, W Miller, E W Myers, and D J Lipman. 1990. "Basic Local Alignment Search Tool." *Journal of Molecular Biology* 215 (3): 403–10. doi:10.1016/S0022-2836(05)80360-2.

Ameur, Adam, Ammar Zaghlool, Jonatan Halvardson, Anna Wetterbom, Ulf Gyllensten, Lucia Cavelier, and Lars Feuk. 2011. "Total RNA Sequencing Reveals Nascent Transcription and Widespread Co-Transcriptional Splicing in the Human Brain." *Nature Structural & Molecular Biology* 18 (12). Nature Publishing Group: 1435–40. doi:10.1038/nsmb.2143.

Anders, Simon, Paul Theodor Pyl, and Wolfgang Huber. 2014. "HTSeq - A Python Framework to Work with High-Throughput Sequencing Data." *Bioinformatics (Oxford, England)* 31 (2): 166–69. doi:10.1093/bioinformatics/btu638.

Andrews, Shea J, and Joseph a Rothnagel. 2014. "Emerging Evidence for Functional Peptides Encoded by Short Open Reading Frames." *Nature Reviews. Genetics* 15 (3): 193–204. doi:10.1038/nrg3520.

Angénieux, Catherine, Blandine Maitre, Anita Eckly, François Lanza, Christian Gachet, and Henri De La Salle. 2016. "Time-Dependent Decay of mRNA and Ribosomal RNA during Platelet Aging and Its Correlation with Translation Activity." *PLoS ONE* 11 (1): 1–24. doi:10.1371/journal.pone.0148064.

Ashwal-Fluss, Reut, Markus Meyer, Nagarjuna Reddy Pamudurti, Andranik Ivanov, Osnat Bartok, Mor Hanan, Naveh Evantal, Sebastian Memczak, Nikolaus Rajewsky, and Sebastian Kadener. 2014. "circRNA Biogenesis Competes with Pre-mRNA Splicing." *Molecular Cell* 56 (1). Elsevier Inc.: 1–12. doi:10.1016/j.molcel.2014.08.019.

Bahn, Jae Hoon, Qing Zhang, Feng Li, Tak Ming Chan, Xianzhi Lin, Yong Kim, David T W Wong, and Xinshu Xiao. 2015. "The Landscape of MicroRNA, Piwi-Interacting RNA, and Circular RNA in Human Saliva." *Clinical Chemistry* 61 (1): 221–30. doi:10.1373/clinchem.2014.230433.

Bailey, Jeffrey A, Zhiping Gu, Royden A Clark, Knut Reinert, Rhea V Samonte, Stuart Schwartz, Mark D Adams, Eugene W Myers, Peter W Li, and Evan E Eichler. 2002. "Recent Segmental Duplications in the Human Genome." *Science (New York, N.Y.)* 297 (5583). United States: 1003–7. doi:10.1126/science.1072047.

Bailleul, Bernard. 1996. "During in Vivo Maturation of Eukaryotic Nuclear mRNA , Splicing Yields Excised Exon Circles." *Nucleic Acids Research* 24 (6): 1015–19. doi:10.1093/nar/24.6.1015.

Banfai, B. 2012. "Long Noncoding RNAs Are Rarely Translated in Two Human Cell Lines." *Genome Res.*

Barrett, Steven P, Peter L Wang, and Julia Salzman. 2015. “Circular RNA Biogenesis Can Proceed through an Exon-Containing Lariat Precursor.” *eLife* 4: e07540. doi:10.7554/eLife.07540.

Bentley, David L. 2014. “Coupling mRNA Processing with Transcription in Time and Space.” *Nature Reviews. Genetics* 15 (3). Nature Publishing Group: 163–75. doi:10.1038/nrg3662.

Berget, S M, C Moore, and P a Sharp. 1977. “Spliced Segments at the 5’ Terminus of Adenovirus 2 Late mRNA.” *Proceedings of the National Academy of Sciences of the United States of America* 76 (8): 3171–75.

Bergmann, Jan H, Jingjing Li, Mélanie A Eckersley-maslin, Frank Rigo, Susan M Freier, and David L Spector. 2015. “Regulation of the ESC Transcriptome by Nuclear Long Noncoding RNAs,” 1336–46. doi:10.1101/gr.189027.114.

Best, Myron G., Nik Sol, Irsan Kooi, Jihane Tannous, Bart A. Westerman, François Rustenburg, Pepijn Schellen, et al. 2015. “RNA-Seq of Tumor-Educated Platelets Enables Blood-Based Pan-Cancer, Multiclass, and Molecular Pathway Cancer Diagnostics.” *Cancer Cell* 28 (5): 666–76. doi:10.1016/j.ccell.2015.09.018.

Birney, Ewan, John a Stamatoyannopoulos, Anindya Dutta, Roderic Guigó, Thomas R Gingeras, Elliott H Margulies, Zhiping Weng, et al. 2007. “Identification and Analysis of Functional Elements in 1% of the Human Genome by the ENCODE Pilot Project.” *Nature* 447 (7146): 799–816. doi:10.1038/nature05874.

Bodzioch, Marek, Katarzyna Lapicka-Bodzioch, Barbara Zapala, Wojciech Kamysz, Beata Kiec-Wilk, and Aldona Dembinska-Kiec. 2009. “Evidence for Potential Functionality of Nuclearly-Encoded Humanin Isoforms.” *Genomics* 94 (4). Elsevier Inc.: 247–56. doi:10.1016/j.ygeno.2009.05.006.

Boland, Michael J., Kristopher L. Nazor, and Jeanne F. Loring. 2014. “Epigenetic Regulation of Pluripotency and Differentiation.” *Circulation Research* 115 (2): 311–24. doi:10.1161/CIRCRESAHA.115.301517.

Bray, Paul F, Steven E McKenzie, Leonard C Edelstein, Srikanth Nagalla, Kathleen Delgrosso, Adam Ertel, Joan Kupper, et al. 2013. “The Complex Transcriptional Landscape of the Anucleate Human Platelet.” *BMC Genomics* 14 (1). BMC Genomics: 1. doi:10.1186/1471-2164-14-1.

Burd, Christin E, William R Jeck, Yan Liu, Hanna K Sanoff, Zefeng Wang, and Norman E Sharpless. 2010. “Expression of Linear and Novel Circular Forms of an INK4/ARF-Associated Non-Coding RNA Correlates with Atherosclerosis Risk.” *PLoS Genetics* 6 (12): e1001233. doi:10.1371/journal.pgen.1001233.

Burkhart, Julia M, Marc Vaudel, Stepan Gambaryan, Sonja Radau, Ulrich Walter, Lennart Martens, Albert Sickmann, and P Zahedi. 2012. “The First Comprehensive and Quantitative Analysis of Human Platelet Protein Composition Allows the Comparative Analysis of Structural and Functional Pathways.” *E-Blood* 120 (15): 73–82. doi:10.1182/blood-2012-04-416594.

Caldas, Carlos, Chi W. So, Angus MacGregor, Anthony M. Ford, Bernadette McDonald, Li C. Chan, and Leanne M. Wiedemann. 1998. “Exon Scrambling of MLL Transcripts Occur Commonly and Mimic Partial Genomic Duplication of the Gene.” *Gene* 208 (2): 167–76. doi:10.1016/S0378-1119(97)00640-9.

Camussi, Giovanni, Maria C Deregibus, Stefania Bruno, Vincenzo Cantaluppi, and Luigi Biancone. 2010. “Exosomes/microvesicles as a Mechanism of Cell-to-Cell Communication.” *Kidney International* 78 (9). United States: 838–48. doi:10.1038/ki.2010.278.

Capel, Blanche, Amanda Swain, Silvia Nicolis, Adam Hacker, Michael Walter, Peter Koopman, Peter Goodfellow, and Robin Lovell-Badge. 1993. “Circular Transcripts of the Testis-Determining Gene Sry in Adult Mouse Testis.” *Cell* 73 (5): 1019–30. doi:10.1016/0092-8674(93)90279-Y.

Carrara, Matteo, Marco Beccuti, Federica Cavallo, Susanna Donatelli, Fulvio Lazzarato, Francesca Cordero, and Raffaele a Calogero. 2013. “State of Art Fusion-Finder Algorithms Are Suitable to Detect Transcription-Induced Chimeras in Normal Tissues?” *BMC Bioinformatics* 14 Suppl 7 (Suppl 7). BioMed Central Ltd: S2. doi:10.1186/1471-2105-14-S7-S2.

Caudevilla, C, D Serra, a Miliar, C Codony, G Asins, M Bach, and F G Hegardt. 1998. “Natural Trans-Splicing in Carnitine Octanoyltransferase Pre-mRNAs in Rat Liver.” *Proceedings of the National Academy of Sciences of the United States of America* 95 (21): 12185–90.

Chen, Edward Y, Christopher M Tan, Yan Kou, Qiaonan Duan, Zichen Wang, Gabriela Vaz Meirelles, Neil R Clark, and Avi Ma’ayan. 2013. “Enrichr: Interactive and Collaborative HTML5 Gene List Enrichment Analysis Tool.” *BMC Bioinformatics* 14 (1): 128. doi:10.1186/1471-2105-14-128.

Chen, Iju, Chia Ying Chen, and Trees Juen Chuang. 2015. “Biogenesis, Identification, and Function of Exonic Circular RNAs.” *Wiley Interdisciplinary Reviews: RNA* 6 (5): 563–79. doi:10.1002/wrna.1294.

Chen, Jianjun, Miao Sun, W. James Kent, Xiaoqiu Huang, Hanqing Xie, Wenquan Wang, Guolin Zhou, Run Zhang Shi, and Janet D. Rowley. 2004. “Over 20% of Human Transcripts Might Form Sense-Antisense Pairs.” *Nucleic Acids Research* 32 (16): 4812–20. doi:10.1093/nar/gkh818.

Chen, Ken, John W Wallis, Cyriac Kandoth, Joelle M Kalicki-Veizer, Karen L Mungall, Andrew J Mungall, Steven J Jones, et al. 2012. “BreakFusion: Targeted Assembly-Based Identification of Gene Fusions in Whole Transcriptome Paired-End Sequencing Data.” *Bioinformatics (Oxford, England)* 28 (14): 1923–24. doi:10.1093/bioinformatics/bts272.

Chen, Liang. 2013. “Characterization and Comparison of Human Nuclear and Cytosolic Editomes.” *Proceedings of the National Academy of Sciences of the United States of America* 110: E2741–47. doi:10.1073/pnas.1218884110.

Chen, Taiping, and Sharon Y R Dent. 2014. “Chromatin Modifiers and Remodellers: Regulators of Cellular Differentiation.” *Nature Reviews. Genetics* 15 (2). Nature Publishing Group: 93–106. doi:10.1038/nrg3607.

Cheng, Ee-chun, and Haifan Lin. 2013. “Repressing the Repressor: A lincRNA as a MicroRNA Sponge in Embryonic Stem Cell Self-Renewal.” *Developmental Cell* 25 (1). United States: 1–2. doi:10.1016/j.devcel.2013.03.020.

Choi, Jung Kyoon. 2010. “Contrasting Chromatin Organization of CpG Islands and Exons in the Human Genome.” *Genome Biology* 11 (7): R70. doi:10.1186/gb-2010-11-7-r70.

Chow, L T, J M Roberts, J B Lewis, and T R Broker. 1977. “A Map of Cytoplasmic RNA Transcripts from Lytic Adenovirus Type 2, Determined by Electron Microscopy of RNA:DNA Hybrids.” *Cell* 11 (4): 819–36. doi:10.1016/0092-8674(77)90294-X.

Choy, Jocelyn Y H, Priscilla L S Boon, Nicolas Bertin, and Melissa J Fullwood. 2015. “A Resource of Ribosomal RNA-Depleted RNA-Seq Data from Different Normal Adult and Fetal Human Tissues.” *Scientific Data* 2: 150063. doi:10.1038/sdata.2015.63.

Clark, Michael B, Anupma Choudhary, Martin A Smith, Ryan J Taft, and John S Mattick. 2013. “The Dark Matter Rises: The Expanding World of Regulatory RNAs.” *Essays in Biochemistry* 54: 1–16. doi:10.1042/bse0540001.

Cocquerelle, C, P Daubersies, M A Majérus, J P Kerckaert, and B Bailleul. 1992. “Splicing with Inverted Order of Exons Occurs Proximal to Large Introns.” *The EMBO Journal* 11 (3): 1095–98.

Cocquerelle, C, B Mascrez, D Hétuin, and B Bailleul. 1993. “Mis-Splicing Yields Circular RNA Molecules.” *The FASEB Journal* 7 (1): 155–60.

Cocquet, Julie, Allen Chong, Guanglan Zhang, and Reiner a Veitia. 2006. “Reverse Transcriptase Template Switching and False Alternative Transcripts.” *Genomics* 88 (1): 127–31. doi:10.1016/j.ygeno.2005.12.013.

Conn, Simon J., Katherine A. Pillman, John Toubia, Vanessa M. Conn, Marika Salmanidis, Caroline A. Phillips, Suraya Roslan, Andreas W. Schreiber, Philip A. Gregory, and Gregory J. Goodall. 2015. "The RNA Binding Protein Quaking Regulates Formation of circRNAs." *Cell* 160 (6). Elsevier Inc.: 1125–34. doi:10.1016/j.cell.2015.02.014.

Cowman, Jonathan, Eimear Dunne, Irene Oglesby, Barry Byrne, Adam Ralph, Bruno Voisin, Sieglinde Müllers, Antonio J. Ricco, and Dermot Kenny. 2015. "Age-Related Changes in Platelet Function Are More Profound in Women than in Men." *Scientific Reports* 5. Nature Publishing Group: 12235. doi:10.1038/srep12235.

Danan, Miri, Schraga Schwartz, Sarit Edelheit, and Rotem Sorek. 2012. "Transcriptome-Wide Discovery of Circular RNAs in Archaea." *Nucleic Acids Research* 40 (7): 3131–42. doi:10.1093/nar/gkr1009.

Davis, Richard E, Cara Hardwick, Paul Tavernier, Scott Hodgson, and Hardeep Singh. 1995. "RNA Trans-Splicing in Flatworms" 270 (37): 21813–19.

de la Mata, Manuel, Celina Lafaille, and Alberto R Kornblihtt. 2010. "First Come, First Served Revisited: Factors Affecting the Same Alternative Splicing Event Have Different Effects on the Relative Rates of Intron Removal." *RNA (New York, N.Y.)* 16 (5): 904–12. doi:10.1261/rna.1993510.

Denis, Melvin M, Neal D Tolley, Michaeline Bunting, Hansjörg Schwertz, Huimiao Jiang, Stephan Lindemann, Christian C Yost, et al. 2005. "Escaping the Nuclear Confines: Signal-Dependent Pre-mRNA Splicing in Anucleate Platelets." *Cell* 122 (3): 379–91. doi:10.1016/j.cell.2005.06.015.

Derrien, Thomas, Rory Johnson, Giovanni Bussotti, Andrea Tanzer, Sarah Djebali, Hagen Tilgner, Gregory Guernec, et al. 2012. "The GENCODE v7 Catalog of Human Long Noncoding RNAs: Analysis of Their Gene Structure, Evolution, and Expression." *Genome Research* 22 (9). United States: 1775–89. doi:10.1101/gr.132159.111.

Deutsch, Varda R, and Aaron Tomer. 2006. "Megakaryocyte Development and Platelet Production." *British Journal of Haematology* 134 (5): 453–66. doi:10.1111/j.1365-2141.2006.06215.x.

Dixon, Richard J, Ian C Eperon, Laurence Hall, and Nilesh J Samani. 2005. "A Genome-Wide Survey Demonstrates Widespread Non-Linear mRNA in Expressed Sequences from Multiple Species." *Nucleic Acids Research* 33 (18). Oxford University Press: 5904–13. doi:doi: 10.1093/nar/gki893.

Dixon, Richard J, Ian C Eperon, and Nilesh J Samani. 2007. "Complementary Intron Sequence Motifs Associated with Human Exon Repetition: A Role for Intragenic, Inter-Transcript Interactions in Gene Expression." *Bioinformatics* 23 (2). Oxford University Press: 150–55. doi:doi: 10.1093/bioinformatics/btl575.

Djebali, Sarah, Julien Lagarde, Philipp Kapranov, Vincent Lacroix, Christelle Borel, Jonathan M Mudge, Cédric Howald, et al. 2012. "Evidence for Transcript Networks Composed of Chimeric RNAs in Human Cells." *PloS One* 7 (1): e28213. doi:10.1371/journal.pone.0028213.

Dobin, Alexander, Carrie a Davis, Felix Schlesinger, Jorg Drenkow, Chris Zaleski, Sonali Jha, Philippe Batut, Mark Chaisson, and Thomas R Gingeras. 2013. "STAR: Ultrafast Universal RNA-Seq Aligner." *Bioinformatics (Oxford, England)* 29 (1): 15–21. doi:10.1093/bioinformatics/bts635.

Dou, Ce, Zhen Cao, Bo Yang, Ning Ding, Tianyong Hou, Fei Luo, Fei Kang, et al. 2016. "Changing Expression Profiles of lncRNAs, mRNAs, circRNAs and miRNAs during Osteoclastogenesis." *Scientific Reports* 6 (February). The Author(s): 21499.

Dubin, Robert a., Manija a. Kazmi, and Harry Ostrer. 1995. "Inverted Repeats Are Necessary for Circularization of the Mouse Testis Sry Transcript." *Gene* 167 (1-2): 245–48. doi:10.1016/0378-1119(95)00639-7.

Eberle, Andrea B., Viktoria Hessle, Roger Helbig, Widad Dantoft, Niclas Gimber, and Neus Visa. 2010. "Splice-Site Mutations Cause Rrp6-Mediated Nuclear Retention of the

Unspliced RNAs and Transcriptional down-Regulation of the Splicing-Defective Genes." *PLoS ONE* 5 (7). doi:10.1371/journal.pone.0011540.
Edelstein, Leonard C, S. E. Mckenzie, C. Shaw, M. A. Holinstat, S. P. Kunapuli, and P. F. Bray. 2013. "MicroRNAs in Platelet Production and Activation." *Journal of Thrombosis and Haemostasis* 11 (SUPPL.1): 340–50. doi:10.1111/jth.12214.
Edelstein, Leonard C, Lukas M Simon, Raúl Teruel Montoya, Michael Holinstat, Edward S Chen, Angela Bergeron, Xianguo Kong, et al. 2013. "Racial Differences in Human Platelet PAR4 Reactivity Reflect Expression of PCTP and miR-376c." *Nature Medicine* 19 (12): 1609–16. doi:10.1038/nm.3385.
Elliott and Ladomery, M., D. 2011. *Molecular Biology of RNA*. Oxford University Press.
Enuka, Yehoshua, Mattia Lauriola, Morris E Feldman, Aldema Sas-Chen, Igor Ulitsky, and Yosef Yarden. 2015. "Circular RNAs Are Long-Lived and Display Only Minimal Early Alterations in Response to a Growth Factor." *Nucleic Acids Research* 44 (3): gkv1367 – . doi:10.1093/nar/gkv1367.
ERCC. 2005. "Proposed Methods for Testing and Selecting the ERCC External RNA Controls." *BMC Genomics* 6. England: 150. doi:10.1186/1471-2164-6-150.
Faghihi, Mohammad Ali, Farzaneh Modarresi, Ahmad M Khalil, Douglas E Wood, Barbara G Sahagan, Todd E Morgan, Caleb E Finch, Georges St. Laurent III, Paul J Kenny, and Claes Wahlestedt. 2008. "Expression of a Noncoding RNA Is Elevated in Alzheimer's Disease and Drives Rapid Feed-Forward Regulation of β-Secretase." *Nature Medicine* 14 (7): 723–30. doi:10.1038/nm1784.
Faghihi, Mohammad, Ming Zhang, Jia Huang, Farzaneh Modarresi, Marcel Van der Brug, Michael Nalls, Mark Cookson, Georges St Laurent, and Claes Wahlestedt. 2010. "Evidence for Natural Antisense Transcript-Mediated Inhibition of microRNA Function." *Genome Biology* 11 (5): R56. doi:doi: 10.1186/gb-2010-11-5-r56.
Fatica, Alessandro, and Irene Bozzoni. 2013. "Long Non-Coding RNAs: New Players in Cell Differentiation and Development." *Nature Reviews. Genetics* 15 (1). Nature Publishing Group: 7–21. doi:10.1038/nrg3606.
Faustino, Nuno André, and Thomas a Cooper. 2003. "Pre-mRNA Splicing and Human Disease." *Genes & Development* 17 (4): 419–37. doi:10.1101/gad.1048803.
Filipowicz, Witold, and Vanda Pogacic. 2002. "Biogenesis of Small Nucleolar Ribonucleoproteins." *Current Opinion in Cell Biology* 14 (3). United States: 319–27.
Frantz, S A, A S Thiara, D Lodwick, L L Ng, I C Eperon, and N J Samani. 1999. "Exon Repetition in mRNA." *Proceedings of the National Academy of Sciences of the United States of America* 96 (10): 5400–5405. doi:10.1073/pnas.96.10.5400.
Frenkel-Morgenstern, Milana, Vincent Lacroix, Iakes Ezkurdia, Yishai Levin, Alexandra Gabashvili, Jaime Prilusky, Angela Del Pozo, et al. 2012. "Chimeras Taking Shape: Potential Functions of Proteins Encoded by Chimeric RNA Transcripts." *Genome Research* 22 (7): 1231–42. doi:10.1101/gr.130062.111.
Friedel, Caroline C, Lars Dölken, Zsolt Ruzsics, Ulrich H Koszinowski, and Ralf Zimmer. 2009. "Conserved Principles of Mammalian Transcriptional Regulation Revealed by RNA Half-Life." *Nucleic Acids Research* 37 (17): e115. doi:10.1093/nar/gkp542.
Gao, Yuan, Jinfeng Wang, and Fangqing Zhao. 2015. "CIRI: An Efficient and Unbiased Algorithm for de Novo Circular RNA Identification." *Genome Biology* 16 (1): 4. doi:10.1186/s13059-014-0571-3.
Garber, Manuel, Manfred G Grabherr, Mitchell Guttman, and Cole Trapnell. 2011. "Computational Methods for Transcriptome Annotation and Quantification Using RNA-Seq." *Nature Methods* 8 (6). United States: 469–77. doi:10.1038/nmeth.1613.
Geiger, Jörg, Julia M Burkhart, Stepan Gambaryan, Ulrich Walter, Albert Sickmann, and René P Zahedi. 2013. "Response: Platelet Transcriptome and Proteome{\textemdash}relation rather than Correlation." *Blood* 121 (26). American Society of Hematology: 5257–58. doi:10.1182/blood-2013-04-493403.

Gelfman, Sahar, Noa Cohen, Ahuvi Yearim, and Gil Ast. 2013. “DNA-Methylation Effect on Cotranscriptional Splicing Is Dependent on GC Architecture of the Exon-Intron Structure.” *Genome Research*. doi:10.1101/gr.143503.112.

Girard, Cyrille, Cindy L. Will, Jianhe Peng, Evgeny M. Makarov, Berthold Kastner, Ira Lemm, Henning Urlaub, Klaus Hartmuth, and Reinhard Lührmann. 2012. “Post-Transcriptional Spliceosomes Are Retained in Nuclear Speckles until Splicing Completion.” *Nature Communications* 3: 994. doi:10.1038/ncomms1998.

Glažar, P., Panagiotis Papavasileiou, and Nikolaus Rajewsky. 2014. “circBase : A Database for Circular RNAs.” *RNA* 20 (11): 1–5. doi:10.1261/rna.043687.113.overview.

Goler-Baron, Vicky, Michael Selitrennik, Oren Barkai, Gal Haimovich, Rona Lotan, and Mordechai Choder. 2008. “Transcription in the Nucleus and mRNA Decay in the Cytoplasm Are Coupled Processes.” *Genes and Development* 22 (15): 2022–27. doi:10.1101/gad.473608.

Gruber, Andreas J, William a Grandy, Piotr J Balwierz, Yoana a Dimitrova, Mikhail Pachkov, Constance Ciaudo, Erik Van Nimwegen, and Mihaela Zavolan. 2014. “Embryonic Stem Cell-Specific microRNAs Contribute to Pluripotency by Inhibiting Regulators of Multiple Differentiation Pathways.” *Nucleic Acids Research* 42 (14): 9313–26. doi:10.1093/nar/gku544.

Guo, Junjie U, Vikram Agarwal, Huili Guo, and David P Bartel. 2014. “Expanded Identification and Characterization of Mammalian Circular RNAs.” *Genome Biology* 15 (7): 409. doi:10.1186/PREACCEPT-1176565312639289.

Guttman, Mitchell, Pamela Russell, Nicholas T Ingolia, Jonathan S Weissman, and Eric S Lander. 2013. “Ribosome Profiling Provides Evidence That Large Noncoding RNAs Do Not Encode Proteins.” *Cell* 154 (1). Elsevier Inc.: 240–51. doi:10.1016/j.cell.2013.06.009.

Ha, Minju, and V Narry Kim. 2014. “Regulation of microRNA Biogenesis.” *Nature Reviews. Molecular Cell Biology* 15 (8). Nature Publishing Group: 509–24. doi:10.1038/nrm3838.

Hacisuleyman, E, L A Goff, C Trapnell, A Williams, J Henao-Mejia, L Sun, P McClanahan, et al. 2014. “Topological Organization of Multichromosomal Regions by the Long Intergenic Noncoding RNA Firre.” *Nat Struct Mol Biol* 21 (2). Nature Publishing Group: 198–206. doi:10.1038/nsmb.2764.

Hacisuleyman, Ezgi, Chinmay J. Shukla, Catherine L. Weiner, and John L. Rinn. 2016. “Function and Evolution of Local Repeats in the Firre Locus.” *Nature Communications* 7. Nature Publishing Group: 11021. doi:10.1038/ncomms11021.

Hansen, Thomas B, Trine I Jensen, Bettina H Clausen, Jesper B Bramsen, Bente Finsen, Christian K Damgaard, and Jørgen Kjems. 2013. “Natural RNA Circles Function as Efficient microRNA Sponges.” *Nature* 495 (7441). Nature Publishing Group, a division of Macmillan Publishers Limited. All Rights Reserved.: 384–88. doi:doi: 10.1038/nature11993.

Hansen, Thomas B, Erik D Wiklund, Jesper B Bramsen, Sune B Villadsen, Aaron L Statham, Susan J Clark, and Jørgen Kjems. 2011. “miRNA-Dependent Gene Silencing Involving Ago2-Mediated Cleavage of a Circular Antisense RNA.” *The EMBO Journal* 30 (21): 4414–22. doi:10.1038/emboj.2011.359.

Hansen, Thomas B., Morten T. Ven??, Christian K. Damgaard, and J??rgen Kjems. 2015. “Comparison of Circular RNA Prediction Tools.” *Nucleic Acids Research* 44 (6): 1–8. doi:10.1093/nar/gkv1458.

Harrow, Jennifer, Adam Frankish, Jose M Gonzalez, and Kelly A Frazer. 2012. “GENCODE : The Reference Human Genome Annotation for The ENCODE Project,” 1760–74. doi:10.1101/gr.135350.111.

Hershfield, V, H W Boyer, C Yanofsky, M A Lovett, and D R Helinski. 1974. “Plasmid ColEl as a Molecular Vehicle for Cloning and Amplification of DNA.” *Proceedings of the National Academy of Sciences of the United States of America* 71 (9). UNITED

STATES: 3455–59.
Hoffmann, Steve, Christian Otto, Gero Doose, Andrea Tanzer, David Langenberger, Sabina Christ, Manfred Kunz, Lesca M Holdt, Daniel Teupser, and Jörg Hackermüller. 2014. “A Multi-Split Mapping Algorithm for Circular RNA , Splicing , Trans-Splicing and Fusion Detection.”
Horiuchi, Takayuki, and Toshiro Aigaki. 2006. “Alternative Trans-Splicing: A Novel Mode of Pre-mRNA Processing.” *Biology of the Cell* 98 (2). Blackwell Publishing Ltd: 135–40. doi:doi: 10.1042/bc20050002.
Houseley, Jonathan, and David Tollervey. 2009. “The Many Pathways of RNA Degradation.” *Cell* 136 (4). Elsevier Inc.: 763–76. doi:10.1016/j.cell.2009.01.019.
———. 2010. “Apparent Non-Canonical Trans-Splicing Is Generated by Reverse Transcriptase in Vitro.” *PloS One* 5 (8): e12271. doi:10.1371/journal.pone.0012271.
Hu, Ganlu, Kevin Huang, Juehua Yu, Sailesh Gopalakrishna-Pillai, Jun Kong, He Xu, Zhenshan Liu, et al. 2012. “Identification of miRNA Signatures during the Differentiation of hESCs into Retinal Pigment Epithelial Cells.” *PLoS ONE* 7 (7). Public Library of Science: e37224. doi:doi: 10.1371/journal.pone.0037224.
Huat, Tee Jong, Amir Ali Khan, Soumya Pati, Zulkifli Mustafa, Jafri Malin Abdullah, and Hasnan Jaafar. 2014. “IGF-1 Enhances Cell Proliferation and Survival during Early Differentiation of Mesenchymal Stem Cells to Neural Progenitor-like Cells.” *BMC Neuroscience* 15 (1): 91. doi:10.1186/1471-2202-15-91.
Hubbard, T, D Barker, E Birney, G Cameron, Y Chen, L Clark, T Cox, et al. 2002. “The Ensembl Genome Database Project.” *Nucleic Acids Research* 30 (1). England: 38–41.
Ingolia, Nicholas T. 2014. “Ribosome Profiling: New Views of Translation, from Single Codons to Genome Scale.” *Nature Reviews. Genetics* 15 (3). Nature Publishing Group: 205–13. doi:10.1038/nrg3645.
Ingolia, Nicholas T, Liana F Lareau, and Jonathan S Weissman. 2011. “Ribosome Profiling of Mouse Embryonic Stem Cells Reveals the Complexity and Dynamics of Mammalian Proteomes.” *Cell* 147 (4). Elsevier Inc.: 789–802. doi:10.1016/j.cell.2011.10.002.
Ivanov, Andranik, Sebastian Memczak, Emanuel Wyler, Francesca Torti, Hagit T. Porath, Marta R. Orejuela, Michael Piechotta, et al. 2014. “Analysis of Intron Sequences Reveals Hallmarks of Circular RNA Biogenesis in Animals.” *Cell Reports*, December. The Authors, 1–8. doi:10.1016/j.celrep.2014.12.019.
Jeck, William R, and Norman E Sharpless. 2014. “Detecting and Characterizing Circular RNAs.” *Nature Biotechnology* 32 (5). Nature Publishing Group: 453–61. doi:10.1038/nbt.2890.
Jeck, William R, Jessica a Sorrentino, Kai Wang, Michael K Slevin, Christin E Burd, Jinze Liu, William F Marzluff, and Norman E Sharpless. 2013. “Circular RNAs Are Abundant, Conserved, and Associated with ALU Repeats.” *RNA (New York, N.Y.)* 19 (2): 141–57. doi:10.1261/rna.035667.112.
Jeggari, Ashwini, Debora S. Marks, and Erik Larsson. 2012. “miRcode: A Map of Putative Microrna Target Sites in the Long Non-Coding Transcriptome.” *Bioinformatics*. doi:10.1093/bioinformatics/bts344.
Ji, Zhe, Ruisheng Song, Aviv Regev, and Kevin Struhl. 2015. “Many lncRNAs, 5’UTRs, and Pseudogenes Are Translated and Some Are Likely to Express Functional Proteins.” *eLife* 4: e08890. doi:10.7554/eLife.08890.
Jiang, Lichun, Felix Schlesinger, Carrie A Davis, Yu Zhang, Renhua Li, Marc Salit, Thomas R Gingeras, and Brian Oliver. 2011. “Synthetic Spike-in Standards for RNA-Seq Experiments.” *Genome Research* 21 (9). United States: 1543–51. doi:10.1101/gr.121095.111.
Kaessmann, Henrik, Nicolas Vinckenbosch, and Manyuan Long. 2009. “RNA-Based Gene Duplication: Mechanistic and Evolutionary Insights.” *Nature Reviews. Genetics* 10 (1): 19–31. doi:10.1038/nrg2487.

Kalyana-Sundaram, Shanker, Chandan Kumar-Sinha, Sunita Shankar, Dan Robinson, Yi-Mi Wu, Xuhong Cao, Irfan Asangani, et al. 2012. “Expressed Pseudogenes in the Transcriptional Landscape of Human Cancers.” *Cell* 149 (7). Cell Press: 1622–34. doi:doi: 10.1016/j.cell.2012.04.041.

Kameyama, Toshiki, Hitoshi Suzuki, and Akila Mayeda. 2012. “Re-Splicing of Mature mRNA in Cancer Cells Promotes Activation of Distant Weak Alternative Splice Sites.” *Nucleic Acids Research* 40 (16): 7896–7906. doi:10.1093/nar/gks520.

Kannan, Kalpana, Liguo Wang, Jianghua Wang, Michael M Ittmann, Wei Li, and Laising Yen. 2011. “Recurrent Chimeric RNAs Enriched in Human Prostate Cancer Identified by Deep Sequencing.” *Proceedings of the National Academy of Sciences of the United States of America* 108 (22): 9172–77. doi:10.1158/1538-7445.AM2011-4973.

Kapranov, P, G St Laurent, T Raz, F Ozsolak, C P Reynolds, P H Sorensen, G Reaman, et al. 2010. “The Majority of Total Nuclear-Encoded Non-Ribosomal RNA in a Human Cell Is ‘Dark Matter’ Un-Annotated RNA.” *BMC Biol* 8 (1). BioMed Central Ltd: 149. doi:1741-7007-8-149 [pii]\r10.1186/1741-7007-8-149.

Kapranov, Philipp, Aarron T Willingham, and Thomas R Gingeras. 2007. “Genome-Wide Transcription and the Implications for Genomic Organization.” *Nature Reviews. Genetics* 8 (6): 413–23. doi:10.1038/nrg2083.

Karginov, Fedor V, and Gregory J Hannon. 2013. “Remodeling of Ago2-mRNA Interactions upon Cellular Stress Reflects miRNA Complementarity and Correlates with Altered Translation Rates.” *Genes & Development* 27 (14): 1624–32. doi:10.1101/gad.215939.113.

Katayama, S, Y Tomaru, T Kasukawa, K Waki, M Nakanishi, M Nakamura, H Nishida, et al. 2005. “Antisense Transcription in the Mammalian Transcriptome.” *Science (New York, N.Y.)* 309 (5740): 1564–66. doi:10.1126/science.1112009.

Kelly, Steven, Chris Greenman, Peter R. Cook, and Argyris Papantonis. 2015. “Exon Skipping Is Correlated with Exon Circularization.” *Journal of Molecular Biology* 427 (15). Elsevier Ltd: 2414–17. doi:10.1016/j.jmb.2015.02.018.

Kent, W J, C W Sugnet, T S Furey, K M Roskin, T H Pringle, a. M Zahler, and a. D Haussler. 2002. “The Human Genome Browser at UCSC.” *Genome Research* 12 (6): 996–1006. doi:10.1101/gr.229102.

Kent, W James. 2002. “BLAT — The BLAST -Like Alignment Tool,” 656–64. doi:10.1101/gr.229202.

Keren, Hadas, Galit Lev-Maor, and Gil Ast. 2010. “Alternative Splicing and Evolution: Diversification, Exon Definition and Function.” *Nat Rev Genet* 11 (5). Nature Publishing Group: 345–55. doi:doi: 10.1038/nrg2776.

Kim, Jin-Soo. 2016. “Genome Editing Comes of Age.” *Nature Protocols* 11 (9): 1573–78. doi:10.1038/nprot.2016.104.

Kim, Min-Sik, Sneha M. Pinto, Derese Getnet, Raja Sekhar Nirujogi, Srikanth S. Manda, Raghothama Chaerkady, Anil K. Madugundu, et al. 2014. “A Draft Map of the Human Proteome.” *Nature* 509 (7502): 575–81. doi:10.1038/nature13302.

Kissopoulou, Antheia, Jon Jonasson, Tomas L Lindahl, and Abdimajid Osman. 2013. “Next Generation Sequencing Analysis of Human Platelet PolyA+ mRNAs and rRNA-Depleted Total RNA.” *PloS One* 8 (12): e81809. doi:10.1371/journal.pone.0081809.

Kodzius, Rimantas, Miki Kojima, Hiromi Nishiyori, Mari Nakamura, Shiro Fukuda, Michihira Tagami, Daisuke Sasaki, et al. 2006. “CAGE: Cap Analysis of Gene Expression.” *Nature Methods* 3 (3): 211–22. doi:10.1038/nmeth0306-211.

Köhler, Alwin, and Ed Hurt. 2007. “Exporting RNA from the Nucleus to the Cytoplasm.” *Nature Reviews Molecular Cell Biology* 8 (September): 761–73. doi:10.1038/nrm2255.

Kramer, Marianne C., Dongming Liang, Deirdre C. Tatomer, Beth Gold, Zachary M. March, Sara Cherry, and Jeremy E. Wilusz. 2015. “Combinatorial Control of Drosophila Circular RNA Expression by Intronic Repeats, hnRNPs, and SR Proteins.” *Genes and*

Development 29 (20): 2168–82. doi:10.1101/gad.270421.115.
Kutmon, Martina, Anders Riutta, Nuno Nunes, Kristina Hanspers, Egon L Willighagen, Anwesha Bohler, Jonathan Mélius, et al. 2016. "WikiPathways: Capturing the Full Diversity of Pathway Knowledge." *Nucleic Acids Research* 44 (D1): D488–94. doi:10.1093/nar/gkv1024.
Lacatena, R M, and G Cesareni. 1981. "Base Pairing of RNA I with Its Complementary Sequence in the Primer Precursor Inhibits ColE1 Replication." *Nature* 294 (5842). ENGLAND: 623–26.
Landry, Patricia, Isabelle Plante, Dominique L Ouellet, Marjorie P Perron, Guy Rousseau, and Patrick Provost. 2009. "Existence of a microRNA Pathway in Anucleate Platelets." *Nature Structural & Molecular Biology* 16 (9). Nature Publishing Group: 961–66. doi:10.1038/nsmb.1651.
Langmead, Ben, and Steven Salzberg. 2012. "Fast Gapped-Read Alignment with Bowtie 2." *Nat Meth* 9 (4). Nature Publishing Group, a division of Macmillan Publishers Limited. All Rights Reserved.: 357–59. doi:doi: 10.1038/nmeth.1923.
Langmead, Ben, Cole Trapnell, Mihai Pop, and Steven L Salzberg. 2009. "Ultrafast and Memory-Efficient Alignment of Short DNA Sequences to the Human Genome." *Genome Biology* 10 (3): R25. doi:10.1186/gb-2009-10-3-r25.
Lasda, Erika, and Roy Parker. 2016. "Circular RNAs Co-Precipitate with Extracellular Vesicles: A Possible Mechanism for Circrna Clearance." *PLoS ONE* 11 (2): 1–11. doi:10.1371/journal.pone.0148407.
Lei, Quan, Cong Li, Zhixiang Zuo, Chunhua Huang, Hanhua Cheng, and Rongjia Zhou. 2016. "Evolutionary Insights into RNA Trans-Splicing in Vertebrates." *Genome Biology and Evolution* 8 (3): 562–77. doi:10.1093/gbe/evw025.
Lenglez, Sandrine, Damien Hermand, and Anabelle Decottignies. 2010. "Genome-Wide Mapping of Nuclear Mitochondrial DNA Sequences Links DNA Replication Origins to Chromosomal Double-Strand Break Formation in Schizosaccharomyces Pombe." *Genome Research* 20 (9). United States: 1250–61. doi:10.1101/gr.104513.109.
Letunic, Ivica, Richard R Copley, and Peer Bork. 2002. "Common Exon Duplication in Animals and Its Role in Alternative Splicing." *Human Molecular Genetics* 11 (13): 1561–67.
Li, Heng, and Richard Durbin. 2009. "Fast and Accurate Short Read Alignment with Burrows–Wheeler Transform." *Bioinformatics*. doi:10.1093/bioinformatics/btp324.
Li, Heng, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, and Richard Durbin. 2009. "The Sequence Alignment/Map Format and SAMtools." *Bioinformatics (Oxford, England)* 25 (16): 2078–79. doi:10.1093/bioinformatics/btp352.
Li, Hui, Jinglan Wang, Gil Mor, and Jeffrey Sklar. 2008. "A Neoplastic Gene Fusion Mimics Trans-Splicing of RNAs in Normal Human Cells." *Science (New York, N.Y.)* 321 (5894). United States: 1357–61. doi:10.1126/science.1156725.
Li, Zhaoyong, Chuan Huang, Chun Bao, Liang Chen, Mei Lin, Xiaolin Wang, Guolin Zhong, et al. 2015. "Exon-Intron Circular RNAs Regulate Transcription in the Nucleus." *Nature Structural & Molecular Biology* 22 (3): 256–64. doi:10.1038/nsmb.2959.
Liang, Dongming, and Jeremy E Wilusz. 2014. "Short Intronic Repeat Sequences Facilitate Circular RNA Production." *Genes & Development*, October. doi:10.1101/gad.251926.114.
Liao, Jo Ling, Juehua Yu, Kevin Huang, Jane Hu, Tanja Diemer, Zhicheng Ma, Tamar Dvash, et al. 2010. "Molecular Signature of Primary Retinal Pigment Epithelium and Stem-Cell-Derived RPE Cells." *Human Molecular Genetics* 19 (21): 4229–38. doi:10.1093/hmg/ddq341.
Lichty, Brian D., Armand Keating, Jeannie Callum, Karen Yee, Ruth Croxford, George Corpus, Bevoline Nwachukwu, Peter Kim, Joyce Guo, and Suzanne Kamel-Reid. 1998.

"Expression of p210 and p190 BCR-ABL due to Alternative Splicing in Chronic Myelogenous Leukaemia." *British Journal of Haematology* 103 (3): 711–15. doi:10.1046/j.1365-2141.1998.01033.x.

Liu, Chenglin, Jinwen Ma, ChungChe Jeff Chang, and Xiaobo Zhou. 2013. "FusionQ: A Novel Approach for Gene Fusion Detection and Quantification from Paired-End RNA-Seq." *BMC Bioinformatics* 14 (1). BMC Bioinformatics: 193. doi:10.1186/1471-2105-14-193.

Liu, Zhenshan, Rongfeng Jiang, Songtao Yuan, Na Wang, Yun Feng, Ganlu Hu, Xianmin Zhu, et al. 2014. "Integrated Analysis of DNA Methylation and RNA Transcriptome during in Vitro Differentiation of Human Pluripotent Stem Cells into Retinal Pigment Epithelial Cells." *PLoS ONE* 9 (3): 1–11. doi:10.1371/journal.pone.0091416.

Londin, Eric R, Eleftheria Hatzimichael, Phillipe Loher, Leonard Edelstein, Chad Shaw, Kathleen Delgrosso, Paolo Fortina, Paul F Bray, Steven E McKenzie, and Isidore Rigoutsos. 2014. "The Human Platelet: Strong Transcriptome Correlations among Individuals Associate Weakly with the Platelet Proteome." *Biology Direct* 9: 3. doi:10.1186/1745-6150-9-3.

Love, Michael I, Wolfgang Huber, and Simon Anders. 2014. "Moderated Estimation of Fold Change and Dispersion for RNA-Seq Data with DESeq2 Moderated Estimation of Fold Change and Dispersion for RNA-Seq Data with DESeq2."

Lu, T T, L L Cui, Y Zhou, C R Zhu, D L Fan, H Gong, Q Zhao, et al. 2015. "Transcriptome-Wide Investigation of Circular RNAs in Rice." *Rna* 21 (12): 2076–87. doi:10.1261/rna.052282.115.

Maniatis, Tom, and Bosiljka Tasic. 2002. "Alternative Pre-mRNA Splicing and Proteome Expansion in Metazoans." *Nature* 418 (6894): 236–43. doi:10.1038/418236a.

Mardis, Elaine R. 2008. "Next-Generation DNA Sequencing Methods." *Annual Review of Genomics and Human Genetics* 9. United States: 387–402. doi:10.1146/annurev.genom.9.081307.164359.

Margulies, Marcel, Michael Egholm, William E Altman, Said Attiya, Joel S Bader, Lisa A Bemben, Jan Berka, et al. 2005. "Genome Sequencing in Microfabricated High-Density Picolitre Reactors." *Nature* 437 (7057). England: 376–80. doi:10.1038/nature03959.

Masuda, Seiji, Rita Das, Hong Cheng, Ed Hurt, Nijsje Dorman, and Robin Reed. 2005. "Recruitment of the Human TREX Complex to mRNA during Splicing," 1512–17. doi:10.1101/gad.1302205.complex.

Maunakea, Alika K, Iouri Chepelev, K R Cui, and K J Zhao. 2013. "Intragenic DNA Methylation Modulates Alternative Splicing by Recruiting MeCP2 to Promote Exon Recognition." *Cell Research* 23 (11). Nature Publishing Group: 1256–69. doi:10.1038/cr.2013.110.

Maurer-Spurej, E, G Pfeiler, N Maurer, H Lindner, O Glatter, and D V Devine. 2001. "Room Temperature Activates Human Blood Platelets." *Laboratory Investigation* 81 (4): 581–92. doi:10.1038/labinvest.3780267.

Mcmanus, C Joel, Gemma E May, Pieter Spealman, and Alan Shteyman. 2014. "Ribosome Profiling Reveals Post-Transcriptional Buffering of Divergent Gene Expression in Yeast," 422–30. doi:10.1101/gr.164996.113.Freely.

Mellough, Carla B., Evelyne Sernagor, Inmaculada Moreno-Gimeno, David H W Steel, and Majlinda Lako. 2012. "Efficient Stage-Specific Differentiation of Human Pluripotent Stem Cells toward Retinal Photoreceptor Cells." *Stem Cells* 30 (4): 673–86. doi:10.1002/stem.1037.

Mellough, Carla, Joseph Collin, Mahmoud Khazim, Kathryn White, Evelyne Sernagor, David Steel, and Majlinda Lako. 2015. "IGF-1 Signaling Plays an Important Role in the Formation of Three-Dimensional Laminated Neural Retina and Other Ocular Structures from Human Embryonic Stem Cells." *Stem Cells*. doi:10.1002/stem.2023.

Memczak, Sebastian, Marvin Jens, Antigoni Elefsinioti, Francesca Torti, Janna Krueger,

Agnieszka Rybak, Luisa Maier, et al. 2013. “Circular RNAs Are a Large Class of Animal RNAs with Regulatory Potency.” *Nature* 495 (7441). Nature Publishing Group: 333–38. doi:10.1038/nature11928.

Merkhofer, Evan C, Peter Hu, and Tracy L Johnson. 2014. “Introduction to Cotranscriptional RNA Splicing.” In *Spliceosomal Pre-mRNA Splicing: Methods and Protocols*, edited by J Klemens Hertel, 83–96. Totowa, NJ: Humana Press. doi:10.1007/978-1-62703-980-2_6.

Mighell, A. J., N. R. Smith, P. A. Robinson, and A. F. Markham. 2000. “Vertebrate Pseudogenes.” *FEBS Letters* 468 (2-3): 109–14. doi:10.1016/S0014-5793(00)01199-6.

Min, Irene M., Joshua J. Waterfall, Leighton J. Core, Robert J. Munroe, John Schimenti, and John T. Lis. 2011. “Regulating RNA Polymerase Pausing and Transcription Elongation in Embryonic Stem Cells.” *Genes and Development* 25 (7): 742–54. doi:10.1101/gad.2005511.

Mondal, Tanmoy, Markus Rasmussen, Gaurav Kumar Pandey, Tanmoy Mondal, Markus Rasmussen, Gaurav Kumar Pandey, Anders Isaksson, and Chandrasekhar Kanduri. 2010. “Characterization of the RNA Content of Chromatin,” 899–907. doi:10.1101/gr.103473.109.

Morrison, Sean J, and Judith Kimble. 2006. “Asymmetric and Symmetric Stem-Cell Divisions in Development and Cancer.” *Nature* 441 (7097): 1068–74. doi:10.1038/nature04956.

Murphy, William J., Kenneth P. Watkins, and Nina Agabian. 1986. “Identification of a Novel Y Branch Structure as an Intermediate in Trypanosome mRNA Processing: Evidence for Trans Splicing.” *Cell* 47 (4): 517–25. doi:10.1016/0092-8674(86)90616-1.

Nacu, Serban, Wenlin Yuan, Zhengyan Kan, Deepali Bhatt, Celina Sanchez Rivers, Jeremy Stinson, Brock a Peters, et al. 2011. “Deep RNA Sequencing Analysis of Readthrough Gene Fusions in Human Prostate Adenocarcinoma and Reference Samples.” *BMC Medical Genomics* 4 (1). BioMed Central Ltd: 11. doi:10.1186/1755-8794-4-11.

Naito, Satoshi, and Hisao Uchida. 1980. “Initiation of DNA Replication in a ColE1-Type Plasmid: Isolation of Mutations in the Ori Region.” *Genetics* 77 (11): 6744–48.

Ng, Shi Yan, Gireesh K. Bogu, BoonSeng Soh, and Lawrence W. Stanton. 2013. “The Long Noncoding RNA RMST Interacts with SOX2 to Regulate Neurogenesis.” *Molecular Cell* 51 (3). Elsevier Inc.: 349–59. doi:10.1016/j.molcel.2013.07.017.

Ng, Shi-Yan, Rory Johnson, and Lawrence W Stanton. 2012. “Human Long Non-Coding RNAs Promote Pluripotency and Neuronal Differentiation by Association with Chromatin Modifiers and Transcription Factors.” *The EMBO Journal* 31 (3). Nature Publishing Group: 522–33. doi:10.1038/emboj.2011.459.

Nigro, J M, K R Cho, E R Fearon, S E Kern, J M Ruppert, J D Oliner, K W Kinzler, and B Vogelstein. 1991. “Scrambled Exons.” *Cell* 64 (3). UNITED STATES: 607–13.

Nishimura, Satoshi, Mika Nagasaki, Shinji Kunishima, Akira Sawaguchi, Asuka Sakata, Hiroyasu Sakaguchi, Tsukasa Ohmori, et al. 2015. “IL-1α Induces Thrombopoiesis through Megakaryocyte Rupture in Response to Acute Platelet Needs.” *Journal of Cell Biology*. doi:10.1083/jcb.201410052.

Nishino, Koichiro, Naoko Hattori, Shun Sato, Yoshikazu Arai, Satoshi Tanaka, Andras Nagy, and Kunio Shiota. 2011. “Non-CpG Methylation Occurs in the Regulatory Region of the Sry Gene.” *Journal of Reproduction and Development* 57 (5): 586–93. doi:10.1262/jrd.11-033A.

Nishino, Koichiro, Naoko Hattori, Satoshi Tanaka, and Kunio Shiota. 2004. “DNA Methylation-Mediated Control of Sry Gene Expression in Mouse Gonadal Development.” *Journal of Biological Chemistry* 279 (21): 22306–13. doi:10.1074/jbc.M309513200.

Odelberg, S J, R B Weiss, a Hata, and R White. 1995. “Template-Switching during DNA Synthesis by Thermus Aquaticus DNA Polymerase I.” *Nucleic Acids Research* 23 (11):

2049–57.
Okano, M, D W Bell, D A Haber, and E Li. 1999. “DNA Methyltransferases Dnmt3a and Dnmt3b Are Essential for de Novo Methylation and Mammalian Development.” *Cell* 99 (3). UNITED STATES: 247–57.
Osman, Abdimajid, Walter E. Hitzler, Adam Ameur, Patrick Provost, and Michael Schubert. 2015. “Differential Expression Analysis by RNA-Seq Reveals Perturbations in the Platelet mRNA Transcriptome Triggered by Pathogen Reduction Systems.” *PLoS ONE* 10 (7): 1–17. doi:10.1371/journal.pone.0133070.
Ota, Hiromitsu, Masayuki Sakurai, Ravi Gupta, Louis Valente, Bjorn-Erik Wulff, Kentaro Ariyoshi, Hisashi Iizasa, Ramana V Davuluri, and Kazuko Nishikura. 2013. “ADAR1 Forms a Complex with Dicer to Promote microRNA Processing and RNA-Induced Gene Silencing.” *Cell* 153 (3). Elsevier Inc.: 575–89. doi:10.1016/j.cell.2013.03.024.
Pei, Baikang, Cristina Sisu, Adam Frankish, Cédric Howald, Lukas Habegger, Xinmeng Jasmine Mu, Rachel Harte, et al. 2012. “The GENCODE Pseudogene Resource.” *Genome Biology* 13 (9). BioMed Central Ltd: R51. doi:10.1186/gb-2012-13-9-r51.
Pelechano, Vicent, and Lars M Steinmetz. 2013. “Gene Regulation by Antisense Transcription.” *Nature Reviews. Genetics* 14 (12). Nature Publishing Group: 880–93. doi:10.1038/nrg3594.
Perriman, R, and M Ares. 1998. “Circular mRNA Can Direct Translation of Extremely Long Repeating-Sequence Proteins in Vivo.” *RNA (New York, N.Y.)* 4 (9): 1047–54. doi:10.1017/S135583829898061X.
Pink, Ryan C, and David R F Carter. 2013. “Pseudogenes as Regulators of Biological Function.” *Essays in Biochemistry* 54: 103–12. doi:10.1042/bse0540103.
Poliseno, Laura, Leonardo Salmena, Jiangwen Zhang, Brett Carver, William J Haveman, and Pier Paolo Pandolfi. 2010. “A Coding-Independent Function of Gene and Pseudogene mRNAs Regulates Tumour Biology.” *Nature* 465 (7301). Nature Publishing Group: 1033–38. doi:10.1038/nature09144.
Pontes, Thaís Brilhante, Caroline De Fátima Aquino Moreira-Nunes, Jersey Heitor Da Silva Maués, Letícia Martins Lamarão, José Alexandre Rodrigues de Lemos, Raquel Carvalho Montenegro, and Rommel Mário Rodriguez Burbano. 2015. “The miRNA Profile of Platelets Stored in a Blood Bank and Its Relation to Cellular Damage from Storage.” *PloS One* 10 (6): e0129399. doi:10.1371/journal.pone.0129399.
Pontier, Daphne B., and Joost Gribnau. 2011. “Xist Regulation and Function eXplored.” *Human Genetics* 130 (2): 223–36. doi:10.1007/s00439-011-1008-7.
Porath, Hagit T, Shai Carmi, and Erez Y Levanon. 2014. “A Genome-Wide Map of Hyper-Edited RNA Reveals Numerous New Sites.” *Nature Communications* 5. Nature Publishing Group: 4726. doi:10.1038/ncomms5726.
Pruitt, Kim D., Tatiana Tatusova, and Donna R. Maglott. 2007. “NCBI Reference Sequences (RefSeq): A Curated Non-Redundant Sequence Database of Genomes, Transcripts and Proteins.” *Nucleic Acids Research* 35 (SUPPL. 1): 501–4. doi:10.1093/nar/gkl842.
Qian, L, M N Vu, M Carter, and M F Wilkinson. 1992. “A Spliced Intron Accumulates as a Lariat in the Nucleus of T Cells.” *Nucleic Acids Research* 20 (20): 5345–50.
Quinlan, Aaron R, and Ira M Hall. 2010. “BEDTools: A Flexible Suite of Utilities for Comparing Genomic Features.” *Bioinformatics (Oxford, England)* 26 (6): 841–42. doi:10.1093/bioinformatics/btq033.
Rearick, David, Ashwin Prakash, Andrew McSweeny, Samuel S. Shepard, Larisa Fedorova, and Alexei Fedorov. 2011. “Critical Association of ncRNA with Introns.” *Nucleic Acids Research* 39 (6): 2357–66. doi:10.1093/nar/gkq1080.
Ren, Ruibao. 2005. “Mechanisms of BCR-ABL in the Pathogenesis of Chronic Myelogenous Leukaemia.” *Nature Reviews. Cancer* 5 (3): 172–83. doi:10.1038/nrc1567.
Ricchetti, Miria, Fredj Tekaia, and Bernard Dujon. 2004. “Continued Colonization of the Human Genome by Mitochondrial DNA.” *PLoS Biology* 2 (9).

doi:10.1371/journal.pbio.0020273.
Rigatti, Roberto, Jian-Hua Jian-Hua Jia, Nilesh J Samani, and Ian C Eperon. 2004. “Exon Repetition: A Major Pathway for Processing mRNA of Some Genes Is Allele-specific.” *Nucleic Acids Research* 32 (2). Oxford University Press: 441–46. doi:doi: 10.1093/nar/gkh197.
Risitano, Antonina, Lea M. Beaulieu, Olga Vitseva, and Jane E. Freedman. 2012. “Platelets and Platelet-like Particles Mediate Intercellular RNA Transfer.” *Blood* 119 (26): 6288–95. doi:10.1182/blood-2011-12-396440.
Robinson, James T, Helga Thorvaldsdottir, Wendy Winckler, Mitchell Guttman, Eric S Lander, Gad Getz, and Jill P Mesirov. 2011. “Integrative Genomics Viewer.” *Nat Biotech* 29 (1). Nature Publishing Group, a division of Macmillan Publishers Limited. All Rights Reserved.: 24–26.
Romero, Gallego Irene, Athma A Pai, Jenny Tung, and Yoav Gilad. 2014. “RNA-Seq: Impact of RNA Degradation on Transcript Quantification.” *BMC Biology* 12 (1): 1–13. doi:10.1186/1741-7007-12-42.
Ruskin, B, and M R Green. 1985. “An RNA Processing Activity That Debranches RNA Lariats.” *Science (New York, N.Y.)* 229 (4709). UNITED STATES: 135–40.
Rybak-Wolf, Agnieszka, Christin Stottmeister, Petar Glažar, Marvin Jens, Natalia Pino, Sebastian Giusti, Mor Hanan, et al. 2015. “Circular RNAs in the Mammalian Brain Are Highly Abundant, Conserved, and Dynamically Expressed.” *Molecular Cell*, April, 1–16. doi:10.1016/j.molcel.2015.03.027.
Salgado, H, G Moreno-Hagelsieb, T F Smith, and J Collado-Vides. 2000. “Operons in Escherichia Coli: Genomic Analyses and Predictions.” *Proceedings of the National Academy of Sciences of the United States of America* 97 (12): 6652–57. doi:10.1073/pnas.110147297.
Salzman, Julia. 2016. “Circular RNA Expression: Its Potential Regulation and Function.” *Trends in Genetics* 32 (5). Elsevier Ltd: 309–16. doi:10.1016/j.tig.2016.03.002.
Salzman, Julia, Raymond E Chen, Mari N Olsen, Peter L Wang, and Patrick O Brown. 2013. “Cell-Type Specific Features of Circular RNA Expression.” *PLoS Genetics* 9 (9): e1003777. doi:10.1371/journal.pgen.1003777.
Salzman, Julia, Charles Gawad, Peter Lincoln Wang, Norman Lacayo, and Patrick O Brown. 2012. “Circular RNAs Are the Predominant Transcript Isoform from Hundreds of Human Genes in Diverse Cell Types.” *PLoS ONE* 7 (2). Public Library of Science: e30733. doi:doi: 10.1371/journal.pone.0030733.
Samonte, Rhea Vallente, and Evan E Eichler. 2002. “Segmental Duplications and the Evolution of the Primate Genome.” *Nature Reviews. Genetics* 3 (1): 65–72. doi:10.1038/nrg705.
Sanchez-Pla, Alex, Ferran Reverter, M Carme Ruiz de Villa, and Manuel Comabella. 2012. “Transcriptomics: mRNA and Alternative Splicing.” *Journal of Neuroimmunology* 248 (1-2). Netherlands: 23–31. doi:10.1016/j.jneuroim.2012.04.008.
Sanger, H L, G Klotz, D Riesner, H J Gross, and A K Kleinschmidt. 1976. “Viroids Are Single-Stranded Covalently Closed Circular RNA Molecules Existing as Highly Base-Paired Rod-like Structures.” *Proceedings of the National Academy of Sciences of the United States of America* 73 (11): 3852–56. doi:10.1073/pnas.73.11.3852.
Sanna, Chaitanya R, Wen-Hsiung Li, and Liqing Zhang. 2008. “Overlapping Genes in the Human and Mouse Genomes.” *BMC Genomics* 9: 169. doi:10.1186/1471-2164-9-169.
Sati, Satish, Vinay Singh Tanwar, K. Anand Kumar, Ashok Patowary, Vaibhav Jain, Sourav Ghosh, Shadab Ahmad, et al. 2012. “High Resolution Methylome Map of Rat Indicates Role of Intragenic DNA Methylation in Identification of Coding Region.” *PLoS ONE* 7 (2): 1–12. doi:10.1371/journal.pone.0031621.
Schoenberg, Daniel R, and Lynne E Maquat. 2012. “Regulation of Cytoplasmic mRNA Decay.” *Nature Reviews. Genetics* 13 (4): 246–59. doi:10.1038/nrg3160.

Schwartz, Schraga, Eran Meshorer, and Gil Ast. 2009. “Chromatin Organization Marks Exon-Intron Structure.” *Nature Structural & Molecular Biology* 16 (9). Nature Publishing Group: 990–95. doi:10.1038/nsmb.1659.

Sen, Rituparno, Suman Ghosal, Shaoli Das, Subrata Balti, and Jayprokas Chakrabarti. 2014. “Competing Endogenous RNA: The Key to Posttranscriptional Regulation.” *TheScientificWorldJournal* 2014. Hindawi Publishing Corporation: 896206. doi:10.1155/2014/896206.

Shao, Xiang, Valery Shepelev, and Alexei Fedorov. 2006. “Bioinformatic Analysis of Exon Repetition, Exon Scrambling and Trans-Splicing in Humans.” *Bioinformatics* 22 (6). Oxford University Press: 692–98. doi:doi: 10.1093/bioinformatics/bti795.

Singer, Meromit, Idit Kosti, Lior Pachter, and Yael Mandel-Gutfreund. 2015. “A Diverse Epigenetic Landscape at Human Exons with Implication for Expression.” *Nucleic Acids Research* 43 (7): 3498–3508. doi:10.1093/nar/gkv153.

Solnick, David. 1985. “Trans Splicing of mRNA Precursors.” *Cell* 42 (1): 157–64. doi:10.1016/S0092-8674(85)80111-2.

Starke, Stefan, Isabelle Jost, Oliver Rossbach, Tim Schneider, Silke Schreiner, Lee-Hsueh Hung, and Albrecht Bindereif. 2014. “Exon Circularization Requires Canonical Splice Signals.” *Cell Reports*, December. The Authors, 1–9. doi:10.1016/j.celrep.2014.12.002.

Surono, a, Y Takeshima, T Wibawa, M Ikezawa, I Nonaka, and M Matsuo. 1999. “Circular Dystrophin RNAs Consisting of Exons That Were Skipped by Alternative Splicing.” *Human Molecular Genetics*. doi:10.1093/hmg/8.3.493.

Tajima, Hirohisa, Takako Niikura, Yuichi Hashimoto, Yuko Ito, Yoshiko Kita, Kenzo Terashita, Kazuto Yamazaki, Atsuo Koto, Sadakazu Aiso, and Ikuo Nishimoto. 2002. “Evidence for in Vivo Production of Humanin Peptide, a Neuroprotective Factor against Alzheimer’s Disease-Related Insults.” *Neuroscience Letters* 324 (3): 227–31. doi:10.1016/S0304-3940(02)00199-4.

Takahara, T, S I Kanazu, S Yanagisawa, and H Akanuma. 2000. “Heterogeneous Sp1 mRNAs in Human HepG2 Cells Include a Product of Homotypic Trans-Splicing.” *The Journal of Biological Chemistry* 275 (48): 38067–72. doi:10.1074/jbc.M002010200.

Takahara, Terunao, Kasahara Daisuke, Mori Daisuke, Shuichi And Yanagisawa, and Akanuma Hiroshi. 2002. “The Trans -Spliced Variants of Sp1 mRNA in Rat” 298: 156–62.

Takahara, Terunao, Bosiljka Tasic, Tom Maniatis, Hiroshi Akanuma, and Shuichi Yanagisawa. 2005. “Delay in Synthesis of the 3’ Splice Site Promotes Trans-Splicing of the Preceding 5' Splice Site.” *Molecular Cell* 18 (2): 245–51. doi:10.1016/j.molcel.2005.03.018.

Tanner, S M, J L Austin, G Leone, L J Rush, C Plass, K Heinonen, K Mrózek, et al. 2001. “BAALC, the Human Member of a Novel Mammalian Neuroectoderm Gene Lineage, Is Implicated in Hematopoiesis and Acute Leukemia.” *Proceedings of the National Academy of Sciences of the United States of America* 98 (24): 13901–6. doi:10.1073/pnas.241525498.

Tilgner, Hagen, David G. Knowles, Rory Johnson, Carrie A. Davis, Sudipto Chakrabortty, Sarah Djebali, Jo??o Curado, Michael Snyder, Thomas R. Gingeras, and Roderic Guig?? 2012. “Deep Sequencing of Subcellular RNA Fractions Shows Splicing to Be Predominantly Co-Transcriptional in the Human Genome but Inefficient for lncRNAs.” *Genome Research* 22 (9): 1616–25. doi:10.1101/gr.134445.111.

Tomecki, Rafal, and Andrzej Dziembowski. 2010. “Novel Endoribonucleases as Central Players in Various Pathways of Eukaryotic RNA Metabolism. TL - 16.” *RNA (New York, N.Y.)* 16 VN - r (9): 1692–1724. doi:10.1261/rna.2237610.

Tran, Nham, and Gyorgy Hutvagner. 2013. “Biogenesis and the Regulation of the Maturation of miRNAs.” *Essays in Biochemistry* 54: 17–28. doi:10.1042/bse0540017.

Trapnell, Cole, Lior Pachter, and Steven L Salzberg. 2009. “TopHat: Discovering Splice

Junctions with RNA-Seq." *Bioinformatics* 25 (9). Oxford University Press: 1105–11. doi:doi: 10.1093/bioinformatics/btp120.
Trapnell, Cole, Adam Roberts, Loyal Goff, Geo Pertea, Daehwan Kim, David R Kelley, Harold Pimentel, Steven L Salzberg, John L Rinn, and Lior Pachter. 2012. "Differential Gene and Transcript Expression Analysis of RNA-Seq Experiments with TopHat and Cufflinks." *Nature Protocols* 7 (3). England: 562–78. doi:10.1038/nprot.2012.016.
Tsumura, Akiko, Tomohiro Hayakawa, Yuichi Kumaki, Shin-ichiro Takebayashi, Morito Sakaue, Chisa Matsuoka, Kunitada Shimotohno, et al. 2006. "Maintenance of Self-Renewal Ability of Mouse Embryonic Stem Cells in the Absence of DNA Methyltransferases Dnmt1, Dnmt3a and Dnmt3b." *Genes to Cells : Devoted to Molecular & Cellular Mechanisms* 11 (7). England: 805–14. doi:10.1111/j.1365-2443.2006.00984.x.
Tuteja, Renu, and Narendra Tuteja. 2004. "Serial Analysis of Gene Expression (SAGE): Unraveling the Bioinformatics Tools." *BioEssays : News and Reviews in Molecular, Cellular and Developmental Biology* 26 (8). United States: 916–22. doi:10.1002/bies.20070.
van Heesch, Sebastiaan, Maarten van Iterson, Jetse Jacobi, Sander Boymans, Paul B Essers, Ewart de Bruijn, Wensi Hao, Alyson W MacInnes, Edwin Cuppen, and Marieke Simonis. 2014. "Extensive Localization of Long Noncoding RNAs to the Cytosol and Mono- and Polyribosomal Complexes." *Genome Biology* 15 (1): R6. doi:10.1186/gb-2014-15-1-r6.
Wagner, E G, and R W Simons. 1994. "Antisense RNA Control in Bacteria, Phages, and Plasmids." *Annual Review of Microbiology* 48. UNITED STATES: 713–42. doi:10.1146/annurev.mi.48.100194.003433.
Wang, Dong, Zhipeng Qu, David L. Adelson, Jian Kang Zhu, and Jeremy N. Timmis. 2014. "Transcription of Nuclear Organellar DNA in a Model Plant System." *Genome Biology and Evolution*. doi:10.1093/gbe/evu111.
Wang, Kai, Darshan Singh, Zheng Zeng, Stephen J Coleman, Yan Huang, Gleb L Savich, Xiaping He, et al. 2010. "MapSplice: Accurate Mapping of RNA-Seq Reads for Splice Junction Discovery." *Nucleic Acids Research* 38 (18). Oxford University Press: e178–e178. doi:doi: 10.1093/nar/gkq622.
Wang, Liguo, Jinfu Nie, Hugues Sicotte, Ying Li, Jeanette E. Eckel-Passow, Surendra Dasari, Peter T. Vedell, et al. 2016. "Measure Transcript Integrity Using RNA-Seq Data." *BMC Bioinformatics* 17 (1). BMC Bioinformatics: 58. doi:10.1186/s12859-016-0922-z.
Wang, Liguo, Shengqin Wang, and Wei Li. 2012. "RSeQC: Quality Control of RNA-Seq Experiments." *Bioinformatics* 28 (16): 2184–85.
Wang, Peter L, Yun Bao, Muh-Ching Yee, Steven P Barrett, Gregory J Hogan, Mari N Olsen, José R Dinneny, Patrick O Brown, and Julia Salzman. 2014. "Circular RNA Is Expressed across the Eukaryotic Tree of Life." *PloS One* 9 (3): e90859. doi:10.1371/journal.pone.0090859.
Wang, Yang, and Zefeng Wang. 2014. "Efficient Backsplicing Produces Translatable Circular mRNAs." *RNA (New York, N.Y.)*, 1–8. doi:10.1261/rna.048272.114.
Wang, Yue, Zhenyu Xu, Junfeng Jiang, Chen Xu, Jiuhong Kang, Lei Xiao, Minjuan Wu, Jun Xiong, Xiaocan Guo, and Houqi Liu. 2013. "Endogenous miRNA Sponge lincRNA-RoR Regulates Oct4, Nanog, and Sox2 in Human Embryonic Stem Cell Self-Renewal." *Developmental Cell* 25 (1). Elsevier Inc.: 69–80. doi:10.1016/j.devcel.2013.03.002.
Wang, Zefeng, Michael E Rolish, Gene Yeo, Vivian Tung, Matthew Mawson, and Christopher B Burge. 2004. "Systematic Identification and Analysis of Exonic Splicing Silencers." *Cell* 119 (6): 831–45. doi:10.1016/j.cell.2004.11.010.
Westholm, Jakub O., and Eric C. Lai. 2011. "Mirtrons: MicroRNA Biogenesis via Splicing." *Biochimie* 93 (11). Elsevier Masson SAS: 1897–1904. doi:10.1016/j.biochi.2011.06.017.
Westholm, Jakub O., Pedro Miura, Sara Olson, Sol Shenker, Brian Joseph, Piero Sanfilippo,

Susan E. Celniker, Brenton R. Graveley, and Eric C. Lai. 2014. “Genome-Wide Analysis of Drosophila Circular RNAs Reveals Their Structural and Sequence Properties and Age-Dependent Neural Accumulation.” *Cell Reports* 9 (5). The Authors: 1966–80. doi:10.1016/j.celrep.2014.10.062.

Weyrich, a S, H Schwertz, L W Kraiss, and G a Zimmerman. 2009. “Protein Synthesis by Platelets: Historical and New Perspectives.” *Journal of Thrombosis and Haemostasis : JTH* 7 (2): 241–46. doi:10.1111/j.1538-7836.2008.03211.x.

Wight, Megan, and Andreas Werner. 2013. “The Functions of Natural Antisense Transcripts.” *Essays in Biochemistry* 54. England: 91–101. doi:10.1042/bse0540091.

Wilusz, Jeffrey. 2015. “Circular RNA and Splicing: Skip Happens.” *Journal of Molecular Biology*, May. Elsevier Ltd, 19–21. doi:10.1016/j.jmb.2015.05.019.

Wilusz, Jeremy E, Hongjae Sunwoo, and David L Spector. 2009. “Long Noncoding RNAs : Functional Surprises from the RNA World Long Noncoding RNAs : Functional Surprises from the RNA World,” 1494–1504. doi:10.1101/gad.1800909.

Wu, Chan-Shuo, Chun-Ying Yu, Ching-Yu Chuang, Michael Hsiao, Cheng-Fu Kao, Hung-Chih Kuo, and Trees-Juen Chuang. 2013. “Integrative Transcriptome Sequencing Identifies Trans-Splicing Events with Important Roles in Human Embryonic Stem Cell Pluripotency.” *Genome Research*. doi:10.1101/gr.159483.113.

Xing, Yi, and Christopher Lee. 2006. “Alternative Splicing and RNA Selection Pressure--Evolutionary Consequences for Eukaryotic Genomes.” *Nature Reviews. Genetics* 7 (7): 499–509. doi:10.1038/nrg1896.

Xu, Jinrui, and Jianzhi Zhang. 2016. “Are Human Translated Pseudogenes Functional’.” *Molecular Biology and Evolution* 33 (3): 755–60. doi:10.1093/molbev/msv268.

Yang, Fan, Xinxian Deng, Wenxiu Ma, Joel B Berletch, Natalia Rabaia, Gengze Wei, James M Moore, et al. 2015. “The lncRNA Firre Anchors the Inactive X Chromosome to the Nucleolus by Binding CTCF and Maintains H3K27me3 Methylation.” *Genome Biology* 16 (1): 52. doi:10.1186/s13059-015-0618-0.

Ye, Chu Yu, Li Chen, Chen Liu, Qian Hao Zhu, and Longjiang Fan. 2015. “Widespread Noncoding Circular RNAs in Plants.” *New Phytologist* 208 (1): 88–95. doi:10.1111/nph.13585.

Yin, Qing-Fei, Li Yang, Yang Zhang, Jian-Feng Xiang, Yue-Wei Wu, Gordon G Carmichael, and Ling-Ling Chen. 2012. “Long Noncoding RNAs with snoRNA Ends.” *Molecular Cell* 48 (2). Elsevier Inc.: 219–30. doi:10.1016/j.molcel.2012.07.033.

You, Xintian, Irena Vlatkovic, Ana Babic, Tristan Will, Irina Epstein, Georgi Tushev, Güney Akbalik, et al. 2015. “Neural Circular RNAs Are Derived from Synaptic Genes and Regulated by Development and Plasticity.” *Nature Neuroscience* 18 (4). Nature Publishing Group: 603–10. doi:10.1038/nn.3975.

Yu, Chun-Ying, Hsiao-Jung Liu, Li-Yuan Hung, Hung-Chih Kuo, and Trees-Juen Chuang. 2014. “Is an Observed Non-Co-Linear RNA Product Spliced in Trans, in Cis or Just in Vitro?” *Nucleic Acids Research* 42 (14): 9410–23. doi:10.1093/nar/gku643.

Yue, Yuan, Yun Yang, Lanzhi Dai, Guozheng Cao, Ran Chen, Weiling Hong, Baoping Liu, et al. 2015. “Long-Range RNA Pairings Contribute to Mutually Exclusive Splicing.” *RNA (New York, N.Y.)*, 1–15. doi:10.1261/rna.053314.115.

Zaphiropoulos, P G. 1997. “Exon Skipping and Circular RNA Formation in Transcripts of the Human Cytochrome P-450 2C18 Gene in Epidermis and of the Rat Androgen Binding Protein Gene in Testis.” *Molecular and Cellular Biology* 17 (6): 2985–93.

Zhang, Jianzhi. 2003. “Evolution by Gene Duplication: An Update.” *Trends in Ecology and Evolution* 18 (6): 292–98. doi:10.1016/S0169-5347(03)00033-8.

Zhang, Xiao-Ou, Qing-Fei Yin, Hai-Bin Wang, Yang Zhang, Tian Chen, Ping Zheng, Xuhua Lu, Ling-Ling Chen, and Li Yang. 2014. “Species-Specific Alternative Splicing Leads to Unique Expression of Sno-lncRNAs.” *BMC Genomics* 15 (1): 287. doi:10.1186/1471-2164-15-287.

Zhang, Yang, Wei Xue, Xiang Li, Jun Zhang, Siye Chen, Jia Lin Zhang, Li Yang, and Ling Ling Chen. 2016. "The Biogenesis of Nascent Circular RNAs." *Cell Reports* 15 (3). The Authors: 611–24. doi:10.1016/j.celrep.2016.03.058.

Zhang, Yang, Xiao-Ou Zhang, Tian Chen, Jian-Feng Xiang, Qing-Fei Yin, Yu-Hang Xing, Shanshan Zhu, Li Yang, and Ling-Ling Chen. 2013. "Circular Intronic Long Noncoding RNAs." *Molecular Cell* 51 (6). Elsevier Inc.: 792–806. doi:10.1016/j.molcel.2013.08.017.

Zhang, Zhengdong D, Adam Frankish, Toby Hunt, Jennifer Harrow, and Mark Gerstein. 2010. "Identification and Analysis of Unitary Pseudogenes: Historic and Contemporary Gene Losses in Humans and Other Primates." *Genome Biology* 11 (3): R26. doi:10.1186/gb-2010-11-3-r26.

Zhao, Shanrong, Wai Ping Fung-Leung, Anton Bittner, Karen Ngo, and Xuejun Liu. 2014. "Comparison of RNA-Seq and Microarray in Transcriptome Profiling of Activated T Cells." *PLoS ONE* 9 (1). doi:10.1371/journal.pone.0078644.

Zheng, Qiupeng, Chunyang Bao, Weijie Guo, Shuyi Li, Jie Chen, Bing Chen, Yanting Luo, et al. 2016. "Circular RNA Profiling Reveals an Abundant circHIPK3 That Regulates Cell Growth by Sponging Multiple miRNAs." *Nature Communications* 7. Nature Publishing Group: 11215. doi:10.1038/ncomms11215.

Zuchner, Stephan, Julia Dallman, Rong Wen, Gary Beecham, Adam Naj, Amjad Farooq, Martin A Kohli, et al. 2011. "Whole-Exome Sequencing Links a Variant in DHDDS to Retinitis Pigmentosa." *American Journal of Human Genetics* 88 (2). United States: 201–6. doi:10.1016/j.ajhg.2011.01.001.

Chapter 9. Appendices

9.1 Supplementary methods

This chapter contains supplementary methods, figures and tables. Lists of identified PTES transcripts, other large tables and scripts can be found in the disc accompanying this thesis.

9.1.1 Primers and Probes

Primer	Sequence
NR_000025.1.1F	CGA GTT CCT TTG GCA GAA GTG
NR_000025.1.1R	CCA CAG AAC ATG GCA CTG AC
NR_002746.1.1F	AAC CGT TCC ATT TTG ATT CTG AGG
NR_002746.1.1R	TTC ATT TGG CAG AAT CAT TAC ATC A
NR_002572.1.1F	TGA AAA CAA CTA CTC TCT GAG CA
NR_002572.1.1R	CCA AAG GTG TCC TAA GAA ATG CC
NR_000019.1.1F	TGA AAG CAG CAC TGG CTG AGA
NR_000019.1.1R	ATC AAA CAA TGT AGG TAG TTG CGG
NR_002585.1.1F	CAG GGC TCC AGT GGG CT
NR_002585.1.1R	TAG GGT GCT CTG GTC CAC AA
NR_002450.1.1F	AAA GCA CAT TTG AAC CCT TTT CCA
NR_002450.1.1R	CAA GGC CGT ACA GCG ATT CC
NM_001164315.10.9F	AGT GAC AAG GAT GAT TCT GTT TCG
NM_001164315.10.9R	CTT CAA GGC CGG TTG TTT CTG
NM_000026.4.3F	CCA GAG TGA TCT CTC GGC TT
NM_000026.4.3R	AAA GCA GGT CAA GTG CAT TTC T
NM_020774.6.2F	TGG TCA TGG AGG ATG GAC TGA
NM_020774.6.2R	TGC ACA CTT CCA TCG AAT GCC
NM_015447.3.2F	AAC GTT CAG TGC CTC GAA AGA
NM_015447.3.2R	CTG CAG ACA CGG CAG TAC A
NM_014976.27.26F	CCC GGA GAC GAA AAG CAA AG
NM_014976.27.26R	TGG ATG ATC GCA GCG ACA AA

Table 9.1. L**ist of snoRNA-PTES primers**. List of primers synthesised for experimental confirmation of mono-exonic PTES from snoRNA genes.

Gene	Structure	Forward Primer	Reverse Primer	Amplicon	Probe (FAM-TAMRA)	Probe	qPCR
PHC3	E4-E5	GGAAGTGTCACACAGCAGTCA	GGGTAATACTGCCGCTGGTA	121	AAGCCGTTCCCAGGCTTCCA	E5	100.08
	E6-E5	TCGTCATCGTCATCTTCCTG		121			93.07
	E7-E5	GGGTAATACTGCCGCTGGTA		123			95.29
SMARCA	E14-15	TCAGACTGGATTCAATAGTCATTCA	CACATGTGTTGCTCCATGTCT	110	GGAGGCTTGTGGATCAGAATCTGAAC	E15	94.17
	E16-E15	TGGGCGAAAGTTCACTTAGAA		169			103.09
UBXN7	E5-E6	TCGGCAAGAACAAGAATTAAGA	TTGCATCTGGCCACACTCT	136	CCGGCCACCCATTGATTTGA	E5	87.95
	E5-E2		CGCTTCAAGCATATGTTTTCC	150			103.62
	E5-E3		TCCTGCTTTTGAGGAATTGG	148			86.95
	E5-E4		CAGGCCGTCGTCTTTTAGG	133			94.62
PNN	E8-E9	GACAAAGCCCCATTTGTTTT	TGCAAATTCGATGCGTCTAC	150	CCTGGAAGAATGTGTCCAGCTACCC	E8	93.58
	E8-E7		CCTGCTTTCTCTCTTCTTCTGC	118			98.03
PPA2	E1-E2	AGCCCTGCTCGCAGAATTAC	AGGAATGCCATTTTCCTCTTT	113	TGGTCACTACATTTCCCCCTTTCATGA	E2	100.08
	E5-E2	AGCTACGCTATGTGGCGAAT		148			99.66
	E6-E2	AAGAGCACGAACTGCTTTGG		140			95.84
PF4	E1-E2	CTGCTGTTCCTGGGGTTG	GGGAGGTGGTCTTCACACAC	106	ACTTGTGGTCGCCTTCGCCA	E1	89.57
PPBP	E1-E2	GCTTCTGTCATTGCTGCTGA	CTTTCCCGATCACTTCCAAA	110	TGCTGAACTCCGCTGCATGTG	E2	101.78
F13A1	E1-E2	CGAGTCCGTTTGAGGAAGTC	TCATCTTCCGCTGCATTAGA	143	CCGCCTTTGGAGGCAGAAGA	E2	103.67
RMST	E12-E6	AAAGTAGAGCTCGTGCTGTGAA	TGGCAACCTATTCCATCACTC	131	TGCCAAGGGGCTAGTTGAGGAA	E12	
RMST	E12-E13	AAAGTAGAGCTCGTGCTGTGAA	AGGATTGAATGCTGCCTACC	131	TGCCAAGGGGCTAGTTGAGGAA	E12	
FIRRE	E10-E5	ATGGGAAGACTTGGTTGTGC	TGTTTGCAAGCCAGGTACAG	296	TATTCTCCTGCCTCAGCCTTCC	E10	
FIRRE	E10-E11	GCATGGATCACTAAGGTCTGTTC	GCACTCCAGCCCTCATAAAG	223	AATGGGAAGACTTGGTTGTGCA	E10	

Table 9.2. **List of qPCR primers and probes.** List of primers and probes used in quantitative PCR assays. See Alhassan et al., 2016 for details of other assays.

9.1.2 RNA concentrations

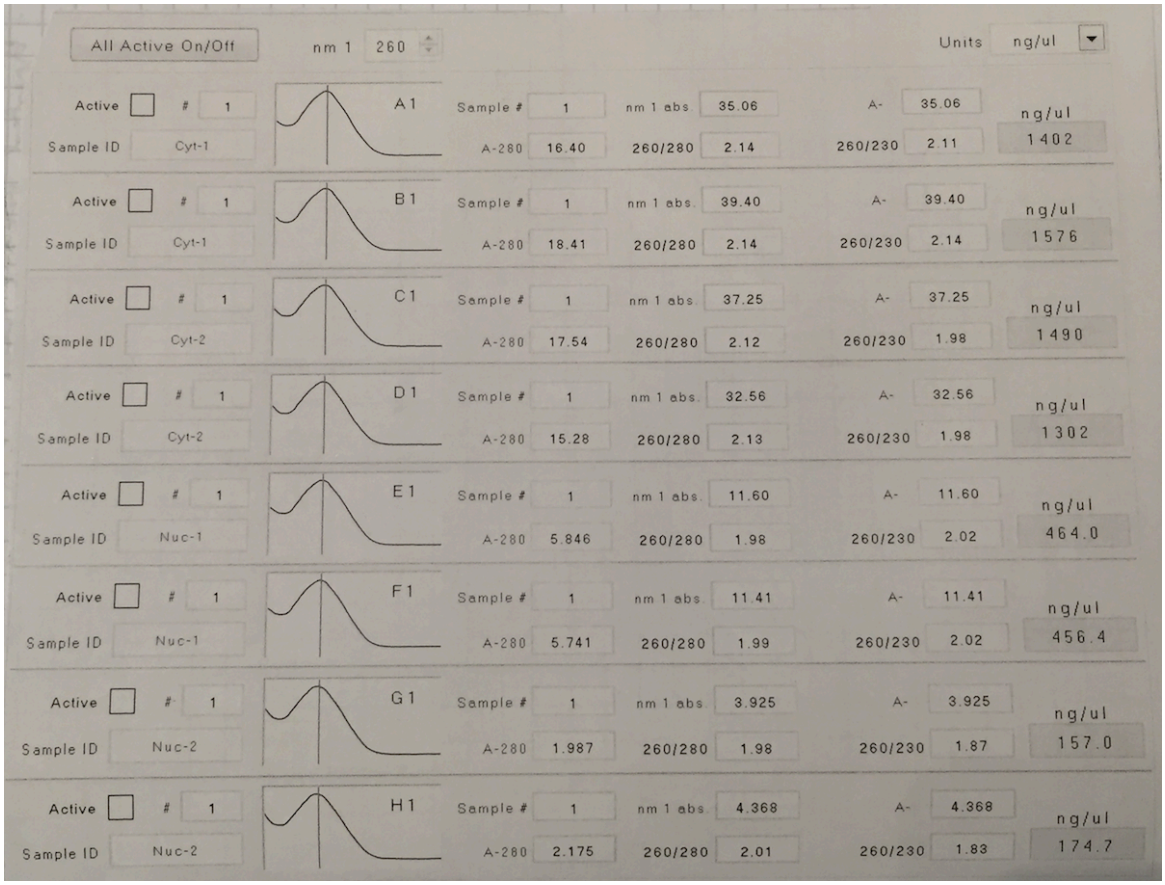


Figure 9.1. **HEK293 and DAOY RNA concentrations**. NanoDrop ND-1000 spectrophotometer results of concentrations of RNA from cytosolic and nuclear fractions of HEK293 (A1, B1, E1 & E2) and DAOY (C1, D1, G1 & H1) cells.

9.1.3 Custom scripts descriptor

See /Appendices/9.1/scripts/

Package	Script	Description / Usage	Inputs	Outputs	Dependencies
ExonicCircRNADecayAnalysis.jar	GenerateCircRNASequence.java ↓ ComputeCircRNASequenceMetrics.java	Generates spliced circRNA sequences by concatenating constituting exonic sequences Used to annotate circRNAs after aligning reads to spliced sequence references of circRNAs. Computes percentile coverages and sequence statistics: GC, miRNA binding sites etc	exons FASTA; circRNAs BED circRNA sequence FASTA; exons BED with miRNA binding sites count; identified circRNAs BED; alingment SAM	circRNA sequences FASTA percentile coverages in plain TEXT file	JAVA 7 JAVA 7; commons-lang3-3%2e2%2e1.jar
SimulateRNAseqData.jar	GenerateMockData.java	Used to generate simulated RNAseq data, at read depths of 2, 5, 10, 25 & 50	number of randomly selected circRNAs; desired read length; number of simulated datasets to generate; canonical junctions BED; circRNAs BED; exons FASTA	multiple FASTQ files for Cjuncs & PTES; list of expected junctions BED	
CotranscriptionalSpliceAssessment.jar	SelectLongestIsoforms.java ↓ GenerateSpliceJunctions.java ↓ ComputeSpliceIndex.java	Used to select longest linear isoform of each multi-exonic genes; expects exon counts in Used to generate sequences for exon-exon and exon-intron splice junctions for all multi-exon Used to post-process alignments to splice junctions and generate co-transcriptional splice rates (CSRs)	refFlat BED from UCSC - both gene ids and transcript ids expected, gene id in column 4 exons BED; sequence segment sixe INTEGER; path to chromosomes FASTA files alingments SAM	selected isoforms BED splice junctions FASTA Text file containing read counts to splice junctions and derived CSRs	
DifferentialDecayAnalysis.jar	ComputeCoverageBedtools; ComputeExonRpkms; NormalizeWithExternalExonsExpression	Used to compute reads coverage across exons, generate exonic read counts and derive expression estimates of PTES/non-PTES exons.	list of circRNAs BED; exons counts from coverageBED in BEDTools 2.15.0; annotated transcripts BED	Text files containing per locus expression estimates of exons internal/external to all circRNAs	BEDTools 2.15.0; JAVA 7
PTESDiscovery.jar		existing transcriptome annotations. Analysis can also be guided by previously identified PTES	see https://sourceforge.net/projects/ptesfinder-v1/ for details of inputs and outputs		
GenomicPTESDiscoverySTAR.jar		Identify PTES junctions, unconstrained by existing transcriptome annotation files	genome reference STAR index; genome FASTA; PID FLOAT; segment size INTEGER; Jspan INTEGER	circRNAs BED	BEDTools; STAR; BOWTIE2; JAVA 7
R	Various scripts	Used for statistical analyses of PTES and mRNA abundance	Count tables: PTES, canonical junctions, exons, genes (HTSeq) and ERCC spike-ins	Lists of differentially expressed PTES, mRNAs	Bioconductor; DESeq; DEXSeq

Table 9.3. **Summary of custom scripts**. List of scripts and software developed for *in silico* analysis performed for this project. Scripts can be found in /Appendices/9.1/scripts folder within the disc accompanying this thesis. Example commands are contained in a text document within the same folder.

9.2 Assessment of PTES identification methods

9.2.1 False positives reported by Memczak et al., 2013

See /Appendices/9.2/List of likely false positive structures in Memczak et al., 2013.xlsx

9.2.2 List of structures identified from human fibroblasts, leukocytes & HEK293 using PTESFinder v.1

See /Appendices/9.2/List of structures identified with PFv1.xlsx

9.2.3 List of structures identified from human fibroblasts using annotation-free PTESFinder

See /Appendices/9.2/Annotation-Free PTESFinder analysis.xlsx

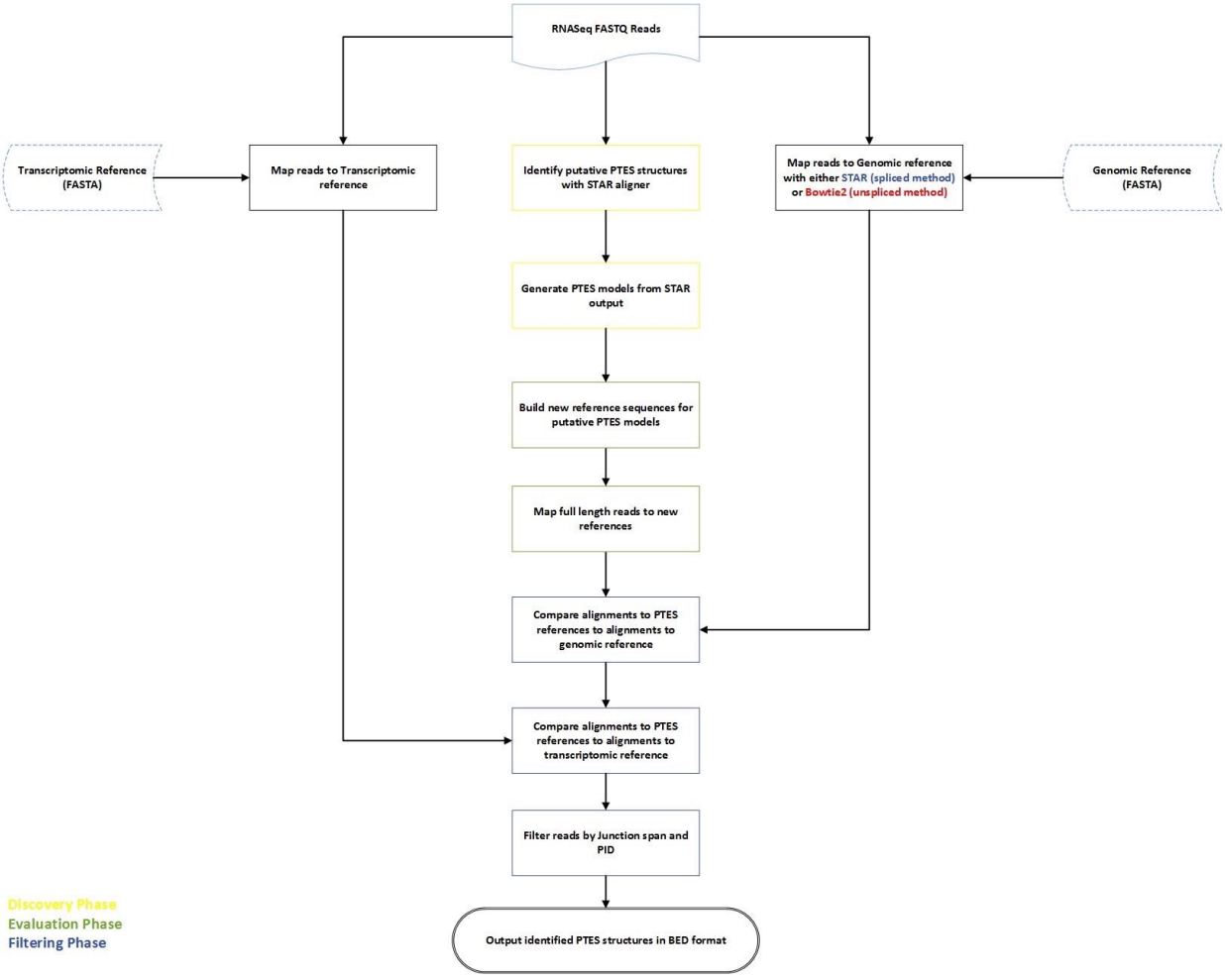


Figure 9.2. **Annotation-Free PTESFinder Workflow**. Pipeline for identifying PTES from RNAseq data, unconstrained by transcriptome annotations. Two versions were developed, one with the STAR aligner, a spliced aligner, as the underlying tool for mapping to the genome (spliced method) and the other with Bowtie2, an unspliced aligner (unspliced method).

9.2.4 Performance test results after varying aligner-specific parameters

See /Appendices/9.2/Varying aligner-specific parameter values.xlsx

9.2.5 Comparison of 5 PTES identification methods

See /Appendices/9.2/Comparisons with published methods.xlsx

9.3 PTES from various cellular compartments

9.3.1 Analysis of PTES and Canonical junctions from nucleus and cytosolic RNA fractions

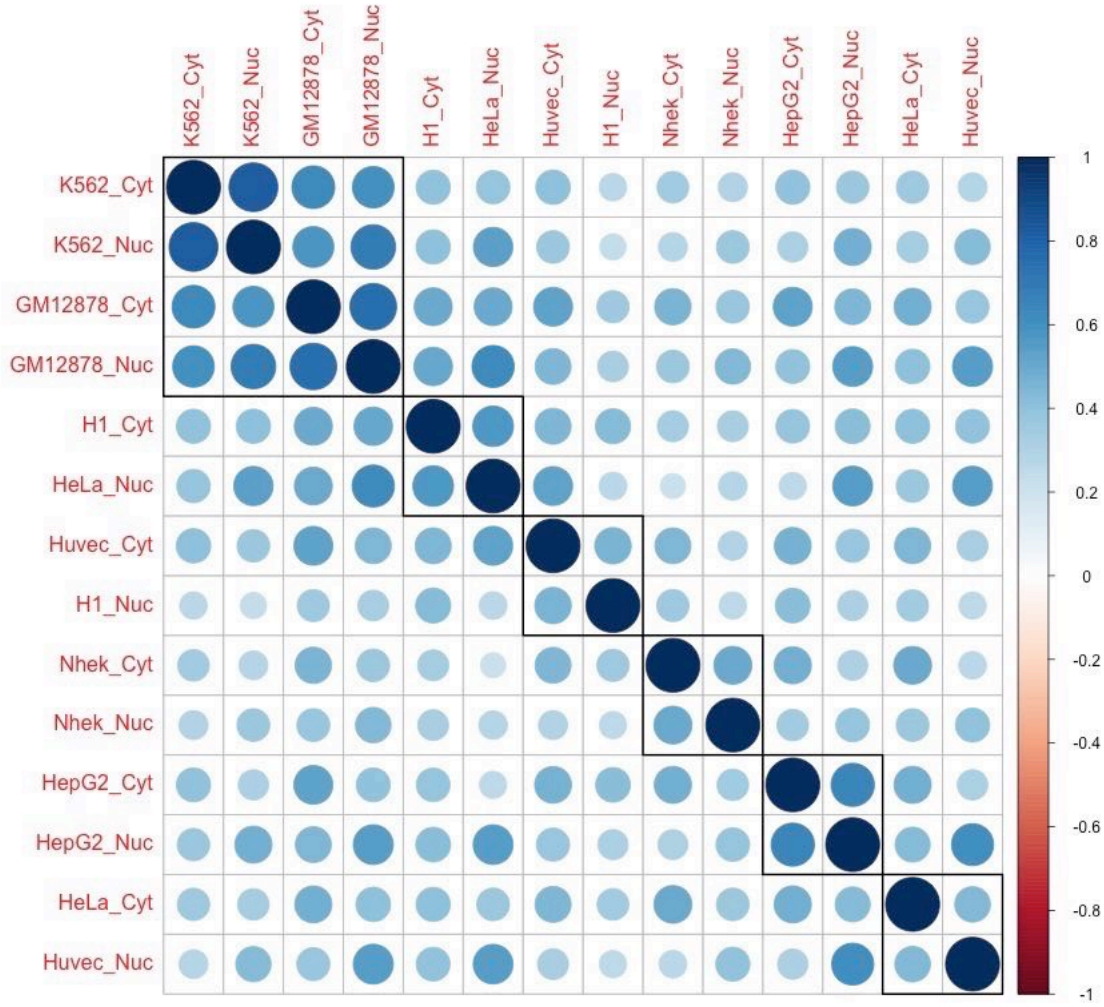


Figure 9.3. **Cluster analysis of PTES in cellular compartments.** Hierarchical clustering and correlational analysis of samples from various 2 cellular compartments of various cell lines based on PTES expression. Samples seemingly cluster by cell lines (instead of cellular compartment), after removing snoRNA-PTES structures.

structures	size	ptes_counts	normalised_counts	snorna_short_read_counts	snorna_long_read_counts	norm_long_read_counts	snorna_abundance	ptes_abundance
NR_002585	134	326	33.898	385	3551	369.239	754.239	4.494%
NR_002746	77	589	61.245	6381	135	14.038	6395.038	0.958%
NR_000025	146	626	65.092	6699	9298	966.821	7665.821	0.849%
NR_002450	72	13	1.352	151	92	9.566	160.566	0.842%
NR_000019	66	106	11.022	1809	7	0.728	1809.728	0.609%
NR_002572	66	235	24.436	4053	0	0.000	4053.000	0.603%
NR_003029	103	17	1.768	1021	188	19.549	1040.549	0.170%
NR_003002	275	29	3.015	87	48690	5062.863	5149.863	0.059%
NR_002963	131	22	2.288	6666	2433	252.987	6918.987	0.033%
NR_002916	109	2	0.208	742	333	34.626	776.626	0.027%
NR_002962	189	12	1.248	4749	19124	1988.544	6737.544	0.019%
NR_004387	330	10	1.040	169	72365	7524.627	7693.627	0.014%
NR_002571	65	123	12.790	107719	0	0.000	107719.000	0.012%
NR_003042	78	9	0.936	7915	136	14.141	7929.141	0.012%
NR_002912	137	17	1.768	18697	2219	230.735	18927.735	0.009%
NR_003010	270	11	1.144	41	150578	15657.338	15698.338	0.007%
NR_003140	76	6	0.624	16488	94	9.774	16497.774	0.004%
NR_002976	132	78	8.111	250507	1084	112.716	250619.716	0.003%
NR_002977	131	16	1.664	53792	99	10.294	53802.294	0.003%
NR_003051	268	170	17.677	483	8316816	864795.621	865278.621	0.002%
NR_002564	74	49	5.095	427650	152	15.805	427665.805	0.001%
NR_002987	130	19	1.976	287921	731	76.011	287997.011	0.001%

Table 9.4. **Analysis of snoRNA-PTES in GM12878.** List of snoRNA-PTES structures identified from GM12878. Read counts of parental genes were extracted from small and long reads libraries, and used to derive abundance ratios. Structures with the highest abundance ratios were selected for downstream analyses. Analysis was performed by Dr. Jackson (IGM, Newcastle University).

9.3.2 Lists of PTES identified from 4 cellular compartments

See /Appendices/9.3/List of PTES transcripts identified from various cellular compartments.xlsx

9.3.3 Genomic alignments (BigWigs)

See /Appendices/9.3/BigWigs

9.3.4 Nucleo-Cytoplasmic enrichment test results

See /Appendices/9.3/List of significantly enriched PTES transcripts.xlsx

9.3.5 Co-transcriptional Splicing Ratios

See /Appendices/9.3/Co-transcriptional Splicing Ratios.xlsx

9.3.6 Terminal exon and PTES junction expression in sucrose-gradient fractions of HEK293

See /Appendices/9.3/List of PTES transcripts identified from various cellular compartments

9.3.7 Sequence analysis of UBAP1.8.7

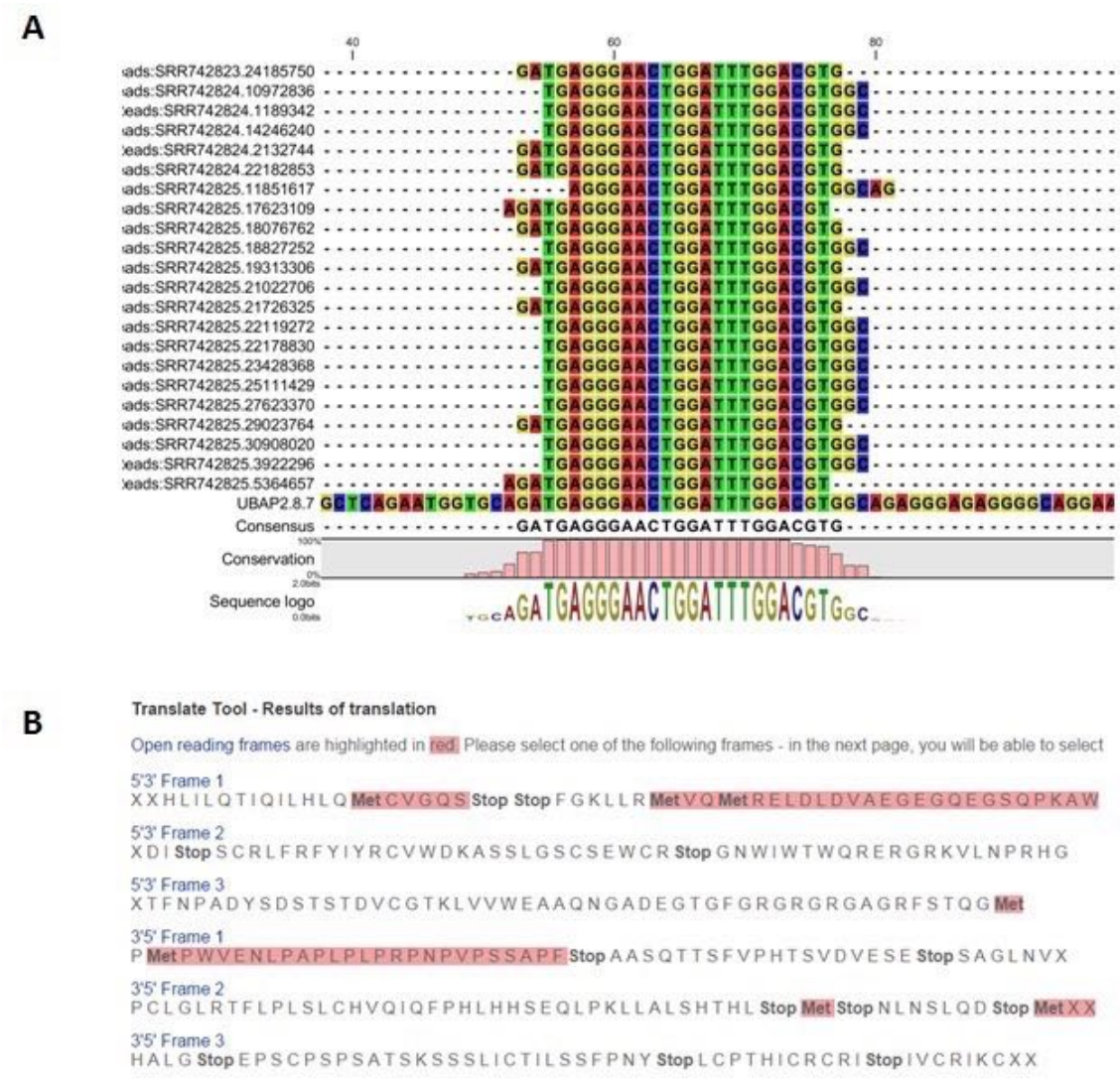


Figure 9.4. **Sequence analysis of UBAP1.8.7**. A) Reads supporting PTES junction between exons 8 and 7 of UBAP2 map across a start codon. B) Six frame translation of full spliced sequence of UBAP.8.7, using Expasy translate tool (http://web.expasy.org/translate/), shows the presence of stop codons upstream of an open reading frame in 5'3' Frame 1.

9.4 PTES in human tissues and anucleate cells

9.4.1 List of PTES identified from tissues and anucleate cells

See /Appendices/9.4/List of PTES identified from anucleate cells and nucleated tissues

9.4.2 Nucleotide composition of PTES in platelets PolyA+ sample

See /Appendices/9.4/Nucleotide composition analysis of PTES identified in Platelets PolyA+ sample.xlsx

9.4.3 Previously unreported PTES identified from platelets

See /Appendices/9.4/Previously unreported PTES identified from Platelets samples.xlsx

9.4.4 Enrichment of PTES exons in Platelets

See /Appendices/9.4/Expression estimates and enrichment of PTES exons in anucleate cells and nucleated tissues.xlsx

9.4.5 List of PTES identified from sub-samples of male platelets sample

See /Appendices/9.4/circrna_decay_analysis/ptes_from_resampled_data/

9.4.6 Percentile coverage of PTES identified in platelets

See /Appendices/9.4/circrna_decay_analysis/Percentile coverage differences.xlsx

9.4.7 Genomic alignments (BigWigs)

See Alhassan et al., (2016)

9.5 PTES in HESC differentiation

9.5.1 List of PTES identified from H9 differentiation series

See /Appendices/9.5/PTES identified from differentiation series.xlsx

9.5.2 RNA editing in PTES flanking introns

See /Appendices/9.5/RNA editing sites in intronic regions flanking identified PTES.xlsx

9.5.3 Differentially expressed PTES in differentiation series

Term	Overlap	P-value	Adjusted P-value	Z-score
EGF/EGFR Signaling Pathway_Homo sapiens_WP437	68/163	2.9E-07	1.1E-04	-2.1E+00
EGFR1 Signaling Pathway_Mus musculus_WP572	65/171	6.7E-06	1.3E-03	-2.2E+00
Insulin Signaling_Mus musculus_WP65	59/154	1.5E-05	1.9E-03	-1.9E+00
BDNF signaling pathway_Homo sapiens_WP2380	55/144	3.1E-05	2.8E-03	-2.1E+00
Insulin Signaling_Homo sapiens_WP481	59/160	3.6E-05	2.8E-03	-1.9E+00
XPodNet - protein-protein interactions in the podocyte expanded by STRING_Mus musculus_WP2309	209/808	7.3E-05	4.7E-03	-2.1E+00
TNF-alpha NF-kB Signaling Pathway_Mus musculus_WP246	60/179	2.7E-04	1.3E-02	-2.1E+00
PluriNetWork_Mus musculus_WP1763	86/284	2.5E-04	1.3E-02	-2.0E+00
TGF-beta Signaling Pathway_Homo sapiens_WP366	47/132	4.2E-04	1.6E-02	-1.8E+00
DNA Replication_Mus musculus_WP150	21/40	4.1E-04	1.6E-02	-1.8E+00
Androgen receptor signaling pathway_Homo sapiens_WP138	35/89	5.4E-04	1.9E-02	-1.6E+00
Retinoblastoma (RB) in Cancer_Homo sapiens_WP2446	35/90	6.3E-04	2.0E-02	-1.6E+00
DNA Replication_Homo sapiens_WP466	21/42	6.7E-04	2.0E-02	-1.6E+00
Interactome of polycomb repressive complex 2 (PRC2) _Homo sapiens_WP2916	12/16	8.0E-04	2.2E-02	-6.2E-01
mRNA processing_Mus musculus_WP310	109/398	9.3E-04	2.4E-02	-1.8E+00
Ectoderm Differentiation_Homo sapiens_WP2858	47/139	1.0E-03	2.5E-02	-1.5E+00
Kit Receptor Signaling Pathway_Mus musculus_WP407	27/67	1.7E-03	3.9E-02	-1.5E+00
Signaling of Hepatocyte Growth Factor Receptor_Homo sapiens_WP313	17/34	2.1E-03	4.4E-02	-1.4E+00
Signaling of Hepatocyte Growth Factor Receptor_Mus musculus_WP193	17/34	2.1E-03	4.4E-02	-1.4E+00
Signaling Pathways in Glioblastoma_Homo sapiens_WP2261	30/83	3.7E-03	7.1E-02	-1.7E+00
Cell Cycle_Homo sapiens_WP179	35/103	4.0E-03	7.5E-02	-1.4E+00
Kit receptor signaling pathway_Homo sapiens_WP304	23/59	5.0E-03	8.9E-02	-1.4E+00
Regulation of Actin Cytoskeleton_Homo sapiens_WP51	46/150	5.5E-03	9.1E-02	-1.6E+00
Integrated Breast Cancer Pathway_Homo sapiens_WP1984	47/155	6.0E-03	9.1E-02	-1.5E+00
Leptin signaling pathway_Homo sapiens_WP2034	27/75	5.9E-03	9.1E-02	-1.5E+00
MicroRNAs in Cardiomyocyte Hypertrophy_Mus musculus_WP1560	28/79	6.1E-03	9.1E-02	-1.3E+00
Regulation of Actin Cytoskeleton_Mus musculus_WP523	45/148	6.8E-03	9.8E-02	-1.4E+00

Table 9.5. **Pathway analysis of PTES genes.** List of pathways enriched by host genes of PTES transcripts reaching statistical significance.

9.5.4 LncRNA PTES producing genes

Chromosome	Start	Stop	Gene	GENCODE ID	Strand	PTES Reads	PTES Transcripts	GENCODE Biotype
chr12	97825430	97958793	RMST	ENST00000538559.2	+	23005	14	processed_transcript
chr5	34164802	34189619	RP11-1023L17.1	ENST00000514048.1	-	8106	7	pseudogene
chr6	26845699	26892992	GUSBP2	ENST00000463434.1	-	7430	4	pseudogene
chrX	130836677	130964671	FIRRE	ENST00000427391.1	-	2857	20	lincRNA
chr2	148656969	148660525	AC009480.3	ENST00000402410.2	-	1954	6	antisense
chr18	9102733	9182522	RP11-21J18.1	ENST00000579126.1	+	1707	5	processed_transcript
chr5	68935285	68975086	GUSBP3	ENST00000513408.2	-	1363	2	pseudogene
chr1	243708833	243711631	RP11-269F20.1	ENST00000439849.1	+	1117	13	antisense
chr14	102502983	102519191	RP11-1017G21.4	ENST00000557242.1	-	994	12	antisense
chr16	68111242	68156174	RP11-67A1.2	ENST00000548144.1	+	964	8	processed_transcript
chr5	129098185	129241354	CTC-575N7.1	ENST00000515569.1	-	709	21	antisense
chr15	72264546	72332606	RP11-390D11.1	ENST00000568391.1	+	657	31	antisense
chr15	63836445	63881254	USP3-AS1	ENST00000559357.1	-	649	19	antisense
chr4	4543930	4712665	STX18-AS1	ENST00000610009.1	+	593	2	antisense
chr6	105279003	105293695	RP11-809N15.2	ENST00000422930.2	-	496	15	sense_overlapping
chr12	53408412	53440743	RP11-983P16.4	ENST00000552905.1	-	475	15	antisense
chr17	45699207	45726786	RP11-580I16.2	ENST00000584391.1	-	462	19	antisense
chr2	61698490	61710061	RP11-355B11.2	ENST00000603028.1	+	443	20	antisense
chr7	74113789	74143187	AC083884.8	ENST00000434256.1	-	438	13	processed_transcript
chr15	63967344	64005957	RP11-317G6.1	ENST00000559303.2	+	425	44	antisense
chr3	114070657	114085891	ZBTB20-AS1	ENST00000496219.1	+	396	2	antisense
chr8	90729655	90769939	RP11-37B2.1	ENST00000504145.1	-	383	2	lincRNA
chr13	76178961	76210130	RP11-173B14.5	ENST00000568735.1	-	376	4	antisense
chr7	72040371	72209725	TYW1B	ENST00000343721.5	-	316	15	polymorphic_pseudogene
chr2	234263219	234301045	AC019221.4	ENST00000442524.1	+	310	1	processed_transcript
chr5	82837295	82877139	VCAN-AS1	ENST00000513899.1	-	308	6	antisense
chr7	128171751	128269512	RP11-274B21.1	ENST00000605862.1	+	302	24	pseudogene

chr10	128811764	128824739	RP11-223P11.2	ENST00000420941.2	-	283	33	antisense
chr17	61271291	61416414	AC037445.1	ENST00000581421.1	-	278	9	antisense
chr3	5198589	5229014	AC026202.3	ENST00000439325.1	-	270	5	antisense
chr20	39726968	39766643	RP1-1J6.2	ENST00000454626.1	-	266	13	antisense
chr6	90539649	90581122	CASP8AP2	ENST00000551025.1	+	260	10	processed_transcript
chr4	146754269	146760732	RP11-181K12.2	ENST00000514334.1	+	252	4	antisense
chr15	85070426	85114026	UBE2Q2P1	ENST00000339094.1	-	249	6	pseudogene
chr16	30296343	30346695	RP11-347C12.2	ENST00000411546.3	-	238	17	pseudogene
chr7	75137074	75157453	PMS2P3	ENST00000418756.1	-	230	6	pseudogene
chr15	23282280	23378228	HERC2P2	ENST00000560464.1	-	219	13	pseudogene
chr6	20756333	20800925	RP3-348I23.2	ENST00000421167.1	-	217	15	antisense
chr9	100013110	100139380	RP11-23J9.4	ENST00000534123.1	+	210	7	processed_transcript
chr5	111563979	111593006	RP11-526F3.1	ENST00000504004.1	+	204	12	antisense
chr7	76610265	76653078	DTX2P1-UPK3BP1-PMS2P11	ENST00000584900.1	+	201	6	processed_transcript
chr15	52877103	52879849	RP11-23N2.4	ENST00000566344.1	+	200	8	antisense
chr15	20588367	20711414	HERC2P3	ENST00000428453.1	-	199	10	pseudogene
chr19	17263425	17278861	CTD-3032J10.2	ENST00000599360.1	-	198	15	antisense
chr15	25427531	25427613	SNHG14	ENST00000365306.1	+	179	8	processed_transcript
chr7	30601177	30612716	AC005154.6	ENST00000582145.1	-	178	1	processed_transcript
chr6	56979708	57037220	RP11-203B9.4	ENST00000416069.2	-	164	16	antisense
chr15	51706293	51791030	RP11-707P17.1	ENST00000561007.1	+	156	9	antisense
chr7	39773230	39832691	LINC00265	ENST00000340510.4	+	154	8	lincRNA
chr11	76666460	76689663	CTD-2547H18.1	ENST00000530190.1	-	148	4	antisense
chr2	243030783	243082789	AC093642.5	ENST00000456398.1	+	146	3	pseudogene
chr7	65874129	65952977	GS1-124K5.2	ENST00000442578.1	-	145	12	pseudogene
chr7	35791465	35840216	AC007551.3	ENST00000437235.3	-	140	2	lincRNA
chr5	54529761	54591029	RP11-506H20.1	ENST00000506435.1	+	139	19	antisense
chr5	68408938	68432221	CTC-498J12.3	ENST00000504129.1	-	130	26	antisense
chr9	2503279	2522019	RP11-125B21.2	ENST00000447278.1	-	125	8	antisense

chr7	77976458	77988775	RPL13AP17	ENST00000450028.1	+	122	2	pseudogene
chr20	34170040	34195484	FER1L4	ENST00000430275.2	-	120	1	pseudogene
chr9	109737113	109865269	RP11-508N12.2	ENST00000439901.1	-	118	3	antisense
chr2	89065384	89106126	ANKRD36BP2	ENST00000393525.3	+	117	4	pseudogene
chr15	23187727	23208417	WHAMMP3	ENST00000400153.2	-	114	8	pseudogene
chr9	123605377	123614881	PSMD5-AS1	ENST00000442982.1	+	113	8	antisense
chr7	64498737	64535091	CCT6P3	ENST00000426828.1	+	110	8	pseudogene
chr2	55541217	55567446	AC012358.8	ENST00000599475.1	+	108	16	antisense
chr7	92119398	92120893	AC007566.10	ENST00000441539.1	+	102	4	antisense
chr22	23995355	24059534	KB-1572G7.2	ENST00000421064.1	-	99	5	processed_transcript
chr1	149239867	149265510	RP11-403I13.4	ENST00000325963.8	+	98	1	lincRNA
chr11	43350297	43380846	RP11-484D2.2	ENST00000526220.1	-	97	4	antisense
chr14	23390249	23396105	PRMT5-AS1	ENST00000590290.1	+	93	3	antisense
chr18	9121262	9136643	RP11-143J12.2	ENST00000582375.1	-	92	3	antisense
chr18	18960246	19030978	RP11-296E23.1	ENST00000584611.1	-	90	5	antisense
chr19	29777917	30016659	CTC-525D6.1	ENST00000582581.1	-	85	1	processed_transcript
chr2	233624855	233632659	AC064852.4	ENST00000427571.1	-	82	4	3prime_overlapping_ncrna
chr16	47333357	47351725	RP11-474B12.1	ENST00000564739.1	+	76	7	antisense
chr6	135622705	135628296	RP3-388E23.2	ENST00000444302.1	+	75	4	antisense
chr2	38257367	38263433	RMDN2-AS1	ENST00000598798.1	-	68	6	antisense
chrX	102024106	102140334	LINC00630	ENST00000440496.1	+	67	5	lincRNA
chr4	119512927	119554884	RP11-384K6.6	ENST00000567913.2	+	66	5	lincRNA
chr1	28906275	28906405	SNHG12	ENST00000384581.1	-	65	2	antisense
chr4	2939321	2948655	NOP14-AS1	ENST00000507702.1	+	63	9	antisense
chr2	160242877	160261143	AC008277.1	ENST00000420020.1	+	62	13	antisense
chr6	107831006	107832640	RP1-67A8.3	ENST00000441532.1	-	58	2	antisense
chr16	11160352	11164959	RP11-66H6.3	ENST00000572828.1	-	57	2	antisense
chr7	74299872	74306687	STAG3L2	ENST00000423186.1	-	55	1	pseudogene
chr5	175774943	175780587	RP11-843P14.2	ENST00000508187.1	+	53	2	antisense
chr6	149276763	149285820	RP11-162J8.2	ENST00000413845.1	-	51	5	antisense

chr22	22657588	22677203	BMS1P20	ENST00000426066.1	+	50	4	processed_transcript
chr15	93425936	93441975	AC013394.2	ENST00000557682.2	+	49	2	processed_transcript
chr7	76875656	76887440	AC073635.5	ENST00000476561.2	-	48	6	antisense
chr8	133850374	133856543	AF230666.2	ENST00000429151.1	-	47	2	antisense
chr6	11173684	11259332	RP3-510L9.1	ENST00000500636.2	+	46	2	antisense
chr3	197880120	197925886	FAM157A	ENST00000437428.2	+	43	4	lincRNA
chr22	43434590	43448372	AL022476.2	ENST00000443063.1	+	42	6	antisense
chr4	419223	467918	ABCA11P	ENST00000451020.2	-	41	4	pseudogene
chr11	63405148	63426434	RP11-697H9.2	ENST00000540307.1	+	40	4	antisense
chr18	13419419	13427479	LDLRAD4-AS1	ENST00000588672.1	-	39	2	antisense
chr13	21872277	21878694	MIPEPP3	ENST00000424756.1	+	38	1	pseudogene
chr21	29811666	30047170	AF131217.1	ENST00000433310.2	-	37	1	lincRNA
chr19	36505409	36536874	AC002116.7	ENST00000586962.1	+	35	3	antisense
chr16	27719746	27730097	CTD-2049O4.1	ENST00000563052.1	-	34	4	antisense
chr15	51749482	51752779	RP11-707P17.2	ENST00000559977.1	+	33	1	antisense
chr2	178563217	178588017	AC012499.1	ENST00000450227.1	+	32	2	antisense
chr1	16793930	16819196	CROCCP3	ENST00000263511.4	-	31	3	pseudogene
chr8	64599736	65281115	RP11-32K4.1	ENST00000523191.1	-	30	5	antisense
chr7	16735495	16759523	AC073333.8	ENST00000418907.1	-	29	6	antisense
chr14	80128008	80257606	RP11-242P2.1	ENST00000553322.1	-	27	4	antisense
chr5	94124493	94129304	CTC-484P3.3	ENST00000513849.1	+	26	3	antisense
chr10	73267909	73271630	CDH23-AS1	ENST00000428918.1	-	25	5	antisense
chr10	54060562	54073888	PRKG1-AS1	ENST00000420193.1	-	24	3	antisense
chr9	71437318	71458191	RP11-203L2.4	ENST00000442103.1	-	23	3	antisense
chr2	186600788	186603752	AC007966.1	ENST00000437717.1	-	22	4	antisense
chr3	67705181	67998137	RP11-81N13.1	ENST00000482677.1	+	21	3	lincRNA
chr4	113567880	113569859	MIR302B	ENST00000505215.1	-	20	4	antisense
chr19	21635997	21646674	CTD-2561J22.5	ENST00000599993.1	-	19	1	processed_transcript
chr12	8448581	8549399	LINC00937	ENST00000544461.1	-	18	2	lincRNA
chr7	5702062	5720092	RNF216-IT1	ENST00000443837.1	-	17	4	sense_intronic

chr6	160424322	160428696	AIRN	ENST00000609176.1	-	16	2	antisense
chr10	29698530	29747716	PTCHD3P1	ENST00000438202.1	+	15	1	antisense
chr5	140997980	141006048	AC008781.7	ENST00000422040.2	+	13	2	antisense
chr15	84748919	84782428	EFTUD1P1	ENST00000558187.1	+	12	1	pseudogene
chr2	47441087	47572105	AC073283.4	ENST00000419035.1	-	11	1	lincRNA
chr12	81488153	81519624	RP11-543H12.1	ENST00000547123.1	-	10	1	antisense

9.6 Publications

List of publications resulting from this project:

- Alhasan AA, **Izuogu OG**, Al-Balool HH, et al. *Circular RNA enrichment in platelets is a signature of transcriptome degradation*. Blood. 2016;127(9): e1-e11. doi:10.1182/blood-2015-06-649434.

- **Izuogu OG**, Alhasan AA, Alafghani HM, Santibanez-Koref M, Elliot DJ, Jackson MS. *PTESFinder: a computational method to identify post-transcriptional exon shuffling (PTES) events*. *BMC Bioinformatics*. 2016; 17:31. doi:10.1186/s12859-016-0881-4.