Transcriptomic consequences of RNA processing disruption via a novel CDC-like kinase inhibitor

by

Tyler Funnell

Bachelor of Science, University of Northern British Columbia, 2009

A THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF

Master of Science

in

THE FACULTY OF GRADUATE AND POSTDOCTORAL STUDIES
(Bioinformatics)

The University of British Columbia
(Vancouver)

December 2014

Abstract

RNA splicing is a process by which introns are excised from precursor mRNA. Variations in the segments removed — and the resulting mRNA molecule — may result in gene transcripts with differing and even opposing functions. The mechanisms involved in RNA splicing are tightly regulated, the disruption of which has been implicated in several human diseases including cancer.

This presents the RNA splicing machinery as a potential therapeutic target. However, the effects of systematic splicing modulation through pharmaceutical intervention remain under explored. A thorough understanding of splicing can be investigated through controlled disruption of the molecular machinery.

The Takeda Pharmaceutical Company Limited (Osaka, Japan) has recently developed a novel compound that inhibits the CDC-like family of kinases, which regulate key splicing factors. Although splicing inhibitors have already been published, their effects on the RNA splicing landscape have not been systematically described. The creation of a novel splicing inhibitor presents the opportunity to perform a methodical analysis of transcriptomic response to RNA processing inhibition using modern RNA sequencing and analysis methods.

It is demonstrated, using the Takeda compound, that restricting the function of CDC-like kinases perturbs RNA splicing in both malignant and normal cells in a dose dependent manner. Post-treatment changes in splicing patterns revealed that these changes are mainly due to inefficient recognition of RNA splice sites. Splicing factors were among the earliest responders to treatment, indicating splicing autoregulatory mechanisms are sensitive to changes in splicing efficiency. Downstream effects were seen as dose-dependent changes in gene expression regulation, and down-regulated genes were enriched for splicing factors. Treatment also resulted in increased generation of conjoined gene transcripts — RNA molecules transcribed from at least two different genes, likely caused by transcriptional read-through. This revelation points to a previously undescribed role for CDC-like kinases in RNA processing.

Preface

This thesis is based on the CLK inhibitor project conducted through a collaboration between the BCCRC and Takeda Pharmaceutical Company. Experimental design and research direction by Sam Aparicio with Osamu Nakanishi, Atsushi Nakanishi, and Gregg Morin. None of the text in this thesis has been previously published.

Experiments were performed by Arusha Oloumi at the BCCRC or by members of the Shonan Incubation Lab at Takeda Pharmaceutical Company. Identification of the T3 compound by Takeda. Arusha Oloumi performed the CLK siRNA experiments and conjoined gene targeted sequencing.

Informatic analysis of experimental data presented in this thesis are performed by myself. Exceptions include RNA-Seq alignment and processing with MISO, which was performed by Jamie Rosner and Celia Siu. First identification of conjoined genes by colleagues at Takeda Pharmaceutical Company. Initial processing of HCT116 T3 treated RNA-Seq libraries with deFuse was performed by Karey Shumansky; All further deFuse processing and analysis was performed by myself with a version of deFuse modified according to Andrew McPherson's advice. ESE motif density analysis was initially performed by Hirokazu Tozaki, but later modified and redone by myself. Differentially spliced gene biological process enrichment analysis was initially performed by Sohrab Shah, and later modified and redone by myself. Selection of clustering method was performed in conjunction with Hiroyoshi Toyoshiba.

Table of Contents

List of Tables

List of Figures

Glossary

A3SS	alternative 3’ splice site
A5SS	alternative 5’ splice site
AFE	alternative first exon
ALE	alternative last exon
AS	alternative splicing
CCS	circular conformation sequencing. see https://github.com/PacificBiosciences/cDNA_primer/wiki/Understanding-PacBio-transcriptome-data
cDNA	complimentary DNA
CG	conjoined gene
CLK	CDC-like kinase
CLR	continuous long read. non CCS PacBio reads
DSE	downstream element
ESE	exonic splicing enhancer
FPKM	Fragments Per Kilobase of exon model per Million mapped reads
mRNA	messenger RNA
MXE	mutually exclusive exon
PSI	percent spliced in
PTC	premature termination codon
RI	retained intron
RNA	ribonucleic acid
RPKM	Reads Per Kilobase of exon model per Million mapped reads
SE	skipped exon
siRNA	small interfering RNA
SMRT	Single Molecule Real Time
snRNP	small nuclear ribonucleoprotein particle

SRPK	SRSF protein kinase
UTR	untranslated region

Chapter 1

Introduction

1.1 Overview

The human genome contains approximately 22,000 protein-coding genes [1]. However, the number of unique protein isoforms is greater than can be explained by the number of genes alone. To reconcile this disparity, we must look at the corpus of gene-protein intermediates: ribonucleic acid (RNA) molecules. Before RNA molecules are ready to be translated into protein peptides, they must undergo a series of modifications to become messenger RNA (mRNA). Splicing is a pre-mRNA processing mechanism that occurs both during and after transcription from genes encoded in the DNA.

A typical eukaryotic gene is primarily composed of exons and introns. Exons are the regions included in the final mRNA product, while introns are the intervening sequences. When pre-mRNA is transcribed from genes, it contains both exons and introns. During splicing, the introns are excised from the RNA molecule and the remaining exons are ligated together, forming mRNA.

Alternative splicing (AS), or the differential inclusion of exons and selection of splice sites, is an important source of proteome diversity in humans. In fact, there are approximately 6 protein coding transcript isoforms per gene on average [2]. Proteins produced from alternatively spliced RNA can have different and even opposing functions. AS can also impact gene expression regulation; In some members of the SR gene family inclusion of a small “poison” exon containing a premature termination codon marks transcripts for decay, reducing transcript abundance within the cell [3].

Alterations in the patterns of AS have been implicated in various human diseases. Approximately 15% of all genetic disease-causing mutations specifically disrupt RNA splicing [4]. In cancer, protein products resulting from aberrant alternative splicing are linked with malignant phenotypes [5, 6]. Modulation of alternative splicing can retard oncogenic activity in tumour cells with relatively low cellular toxicity, suggesting the splicing machinery may be targeted for therapeutic intervention [7]. However, more research is needed to obtain a detailed understanding of the dynamics of splicing regulation and the effects of its suppression before AS-modulating agents can be used (safely) as clinical therapeutics.

RNA sequencing (RNA-Seq) [8] assays allow the precise measurement of nucleotide sequences and quantification of RNA levels. Many methods have been developed that use RNA-Seq data to detect and quantify RNA isoforms with high sensitivity. Recent advances in sequencing technologies, including the Pacific Biosciences (PacBio) [9] RS platform, provide the ability to sequence up to several thousand nucleotides – enough to capture entire transcripts for many genes. Using these long read methodologies, it is possible to verify the existence of specific mRNA splice variants. Current RNA assays provide the ability to study RNA expression and processing mechanisms with unprecedented resolution.

1.2 The Process and Mechanisms of RNA Splicing

RNA splicing is a complex procedure that requires a collaboration of many distinct proteins and ribonucleoprotein particles. For splicing to occur, a subset of these splicing factors assemble onto the mRNA precursor around exon junctions to form the spliceosome complex. Once in place, the spliceosome cleaves the RNA molecule, removing the non-coding intron segment, and ligates the remaining exons together. Recognition and precise definition of exon boundaries involves several *cis*- and *trans*-acting elements that can either promote or inhibit splicing at a candidate exon junction.

1.2.1 Formation of the spliceosome

Formation of the spliceosome involves the cooperative action of five small nuclear ribonucleoprotein particles (snRNPs) in conjunction with many auxiliary proteins. Spliceosomal snRNPs and protein factors recognise and interact with several cis-elements including the $5'$ splice site, branch point, polypyrimidine tract, and $3'$ splice site during assembly of the spliceosome. Assembly proceeds in a step-wise fashion, forming several intermediate complexes before forming the final spliceosome complex [10].

The first pre-spliceosomal complex is formed when the U1 snRNP and splicing factor 1 (SF1) bind to the $5'$ splice site and branch point of an intron, respectively, to form the E' complex. The E' complex is transformed into the E complex with the binding of U2 auxiliary factor (U2AF) to the polypyrimidine tract and $3'$ splice site. The E complex can be converted to the A complex if U2 snRNP is recruited to the pre-mRNA intron through interactions with U2AF, replacing SF1 at the branch site. Recruitment of the U4/U6-U5 tri-snRNP to the A complex generates the B complex. Subsequent extensive rearrangements produce the C spliceosome complex. The C complex catalyzes the next step in the splicing process before disassociating [10].

1.2.2 The role of SR proteins in RNA splicing

Regulation of splicing can occur at many different stages in the process of splice site selection and spliceosome formation. Splicing regulation involves cis-regulatory elements that are categorized into four groups: exonic splicing enhancers (ESE) and silencers (ESS), and intronic splicing enhancers (ISE) and silencers (ISS). ESEs are common, degenerate exonic sequences commonly bound by members of the SR protein family to promote splicing.

SR proteins are characterized by the presence of a C-terminal Serine/Arginine-rich RS domain and at least one N-terminal RNA recognition motif (RRM). Traditional models of SR protein function maintain that the RRM domain mediates interaction between the SR protein and splicing regulatory elements (*e.g.* ESEs) [11], while the RS domain mediates protein-protein interactions with other splicing factors. For example, the RS domain is believed to facilitate the recruitment of U1 snRNP, U2AF, and U2 snRNP to the pre-mRNA substrate [10]. However, studies have also shown that the RS domain contacts the RNA itself during spliceosome formation [12, 13], and that the RRM of SRSF1 is directly involved in recruiting U1 snRNP to the 5′ splice site [14].

1.2.3 The role of SRSF protein kinases and CDC-like kinases in RNA splicing

Serine residues of SR protein RS domains are phosphorylated by members of several protein kinase families, including the CDC-like kinases (CLKs) and the SRSF protein kinases (SRPKs). SRPK-mediated phosphorylation of SR proteins located in the cytoplasm results in their nuclear entry, and concentration in speckles. Subsequent phosphorylation by CLK is necessary for intra-nuclear localization and activation of splicing [15] (see Figure 1.1). Although the exact manner by which this activity regulates splicing is not completely understood, recruitment of spliceosomal components by some SR proteins are thought to occur via phosphorylation-enhanced interactions with the SR protein RS domains [16] (see Figure 1.2a). However, in the case of SRSF1, hyper-phosphorylation of the RS domain promotes the recruitment of U1 snRNP via an RRM-RRM interaction [14] (see Figure 1.2b). Regardless of the precise mechanism, RS domain phosphorylation is a critical step in the formation of the spliceosome.

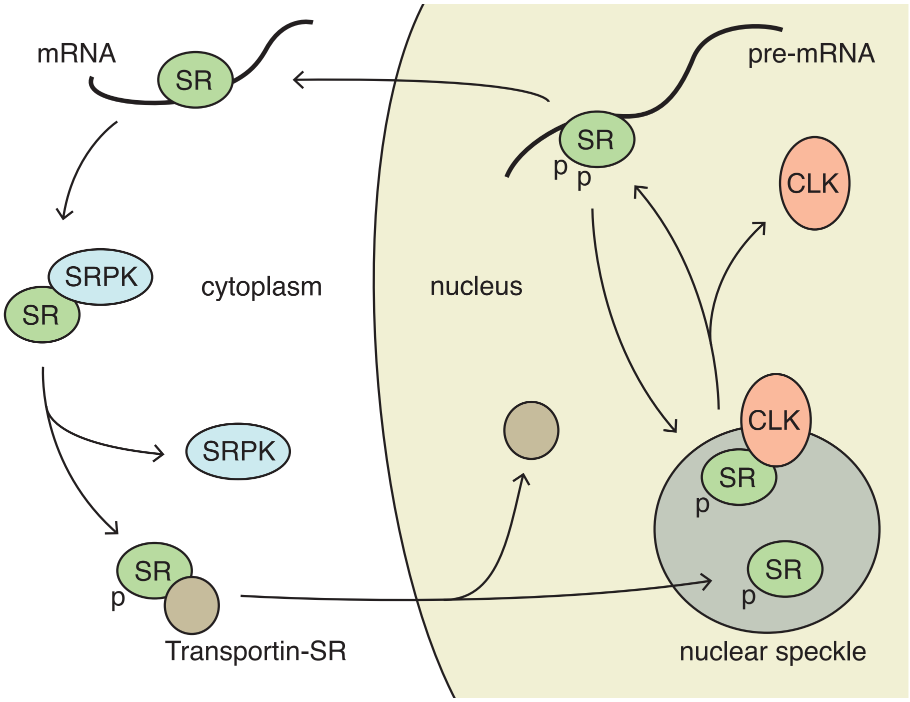

Figure 1.1: Regulation of SR protein cellular localization by phosphorylation. SR proteins in the cytoplasm are phosphorylated by SRPKs which promotes interactions with Transportin-SR and nuclear entry. Within the nucleus, SR proteins tend to aggregate in nuclear speckles until further phosphorylation allows them to dissociate from the speckles and participate in spliceosome formation. Dephosphorylation of SR protein RS domains is necessary for splicing catalysis [17]. Once splicing is complete, SR proteins may either remain associated with the mRNA to facilitate nuclear export and translation, or remain in the nucleus and engage in further splicing reactions.

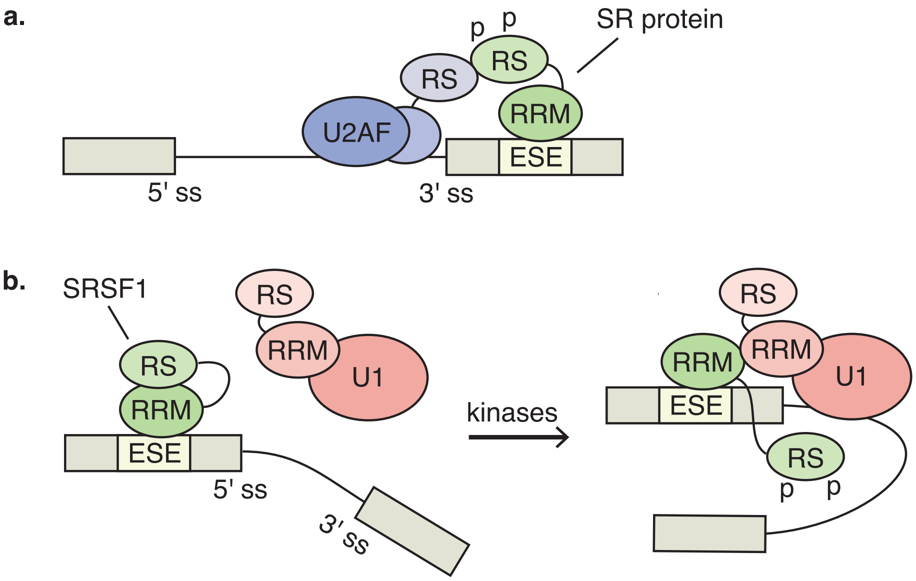

Figure 1.2: Two models for spliceosome recruitment by SR proteins. Phosphorylated RS domains are indicated by the presence of lowercase ‘p’s. **a**, Phosphorylated RS domain mediated recruitment of U2AF via the U2AF 35 kDa subunit’s RS domain. **b**, Recruitment of U1 snRNP by SRSF1. The un- or hypo-phosphorylated RS domain of SRSF1 interacts with a non-RNA-bound interface of the RRM domain. Subsequent phosphorylation of SRSF1’s RS domain disassociates it from the RRM, leaving the RRM open for interaction with the U1 snRNP 70 kDa subunit’s RRM domain. Inspired by figure 6 of [14] and figure 1 of [18].

1.2.4 Alternative splicing

The selection of exon junctions during RNA splicing can be variable. Changes in the set of selected splice sites will impact the structural composition of the final RNA molecule. The exonic structural consequences can be grouped into eight categories of AS events [19] (see Figure 1.3).

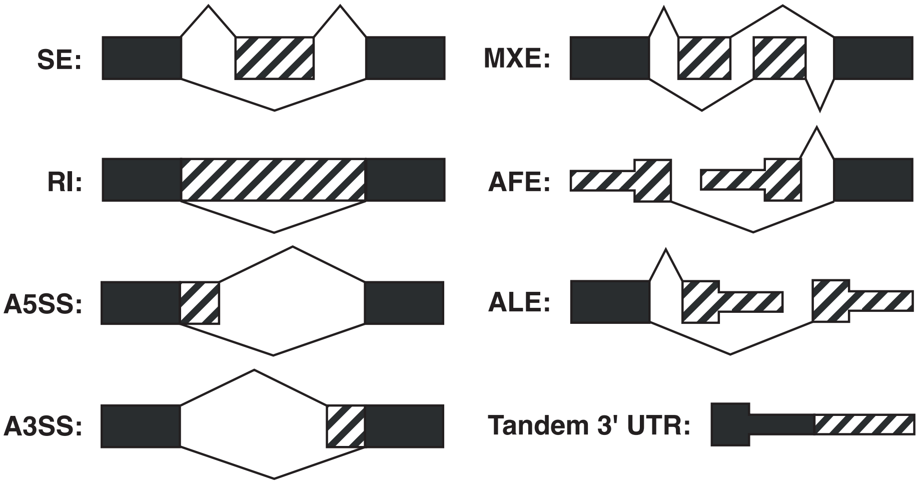

Figure 1.3: Alternative splicing event types. Constitutive exonic regions are solid black. Regions that may be differentially included are striped. Thin black lines represent introns. SE: skipped exon, RI: retained intron, A5SS: alternative 5′ splice site, A3SS: alternative 3′ splice site, MXE: mutually exclusive exons, AFE: alternative first exon, ALE: alternative last exon. Inspired by figure 2 from [19]

Changes in splice site selection can, for example, result in the exclusion of entire exons, as with skipped exon (SE) and mutually exclusive exon (MXE) events; Or, they may cause a shift in the location of an exon's boundaries as with alternative 3' splice site (A3SS) and alternative 5' splice site (A5SS) events [19]. AS events affecting the termini of RNA transcripts (*e.g.* alternative first exon (AFE) and alternative last exon (ALE) events) can result in changes to their untranslated region (UTR) sequences, which can affect transcript stability and localization [20]. This transcriptomic flexibility equips the cell with another regulatory mechanism with which to fine tune gene function.

Various factors play a role in determining the precise locations of splice sites. Recognition of splice sites is regulated in part by the binding of splicing factors (*e.g.* SR proteins) to splicing enhancer and silencer elements within the exonic and surrounding intronic sequences. The relative concentrations and activities of these splicing factors affects the ability of the spliceosome to assemble on exon junctions [10]. Therefore, AS (and overall splicing activity) can be modulated by altering splicing factors' expression, localization, or functional efficacy. For example, disrupting the phosphorylation of SR proteins could negatively impact splicing regulatory programmes.

1.2.5 The role of CLK and SR proteins in non-splicing RNA metabolic processes

The role of SR proteins in RNA splicing and their regulation through CLK-mediated phosphorylation has been established. However, members of the SR protein family are also involved in non-splicing RNA metabolic reactions, including formation of the exon junction complex (EJC) [21], and $3'$ end formation [20, 22]. Disruption of regular SR protein activity may prevent SR proteins from fulfilling their role in other cellular processes.

EJCs assemble upstream of spliced RNA exon-exon junctions and play a number of roles including promotion of mRNA export and translation. However, it is perhaps most well known for its function in the nonsense-mediated mRNA decay pathway; If an mRNA molecule contains a pre-mature stop codon upstream of an EJC, that transcript is marked for degradation. Several SR proteins have been found to interact with the EJC core and may act to stabilize it [21]. Preventing SR proteins from loading on the pre-mRNA substrate or interacting with other proteins may not only reduce levels of splicing, but also broadly inhibit mRNA transport and translation.

For the majority of eukaryotic transcripts, formation of the $3'$ end entails the cleavage of the nascent RNA molecule, followed by the appending of a poly-adenine (poly(A)) tail to the $5'$ cleaved end. The location of cleavage and poly-adenylation is subject to regulation, and at least half of human genes are alternatively poly-adenylated [23]. Alternative poly-adenylation allows for a greater diversity of RNA messages and, consequently, proteins. In this sense, alternative poly-adenylation is similar to alternative splicing.

Generally, recognition of poly(A) sites begins with the binding of cleavage and polyadenylation specificity factor (CPSF) to an A(A/U)UAAA poly(A) signal hexamer in conjunction with the binding of cleavage stimulation factor (CstF) to a U/GU-rich downstream element (DSE). The subsequent steps of cleavage and poly-adenylation are performed by these core proteins along with with a collection of other 3′ processing factors.

CPSF also recognises non-canonical poly(A) signals with reduced efficiency. In these cases, poly(A) site recognition relies on the cooperative action of auxiliary 3′ end processing factors, including cleavage factor I and II (CFIm, CFIIm). CFIm recognizes a UGUA signal upstream of the poly(A) site and recruits CPSF to the unprocessed RNA transcript [24].

CFIm is composed of a 25 kDa subunit and a large subunit of either 59, 68, or 72 kDa. The structures of the 59 and 68 kDa subunit proteins are similar to SR proteins due to their inclusion of both an RNA-binding domain, and an RS-like alternating charge domain. CFIm has been demonstrated to interact with SR proteins [25]. Interactions between SR proteins and CFIm may work to promote binding of CFIm to the RNA substrate and recognition of non-canonical poly(A) sites [24].

The phosphorylation status of CFIm can affect 3′ end formation efficiency. Dephosphorylation of CFIm using Serine/Threonine phosphatases results in the loss of 3′ transcript end cleavage activity in HeLa cell nuclear extract [26]. Dephosphorylation of CPSF and CstF do not produce the same effect. Although the kinase(s) responsible for phosphorylating CFIm are not known, CLKs may be responsible for phosphorylating the CFIm RS-like domain.

The loading of CstF onto the poly(A) site U/GU-rich DSE is an early and essential step of the 3′ cleavage and polyadenylation process. Like CFIm binding of UGUA elements, CstF binding to DSEs promotes selection of poly(A) sites with non-canonical poly(A) signals [27]. SR proteins can affect CstF binding affinity to regulate alternative 3′ end processing. SRSF3 recognition of splicing enhancer signals of the calcitonin/calcitonin gene-related peptide (CT/CGRP) gene promotes recruitment of CstF to the poly(A) site at exon 4 [22]. SRSF3's influence on CstF binding to Poly-A sites may involve CFIm as CFIm binds early in the 3′ end cleavage reaction, and promotes the recruitment of other core 3′ end processing factors.

RS domain phosphorylation status is known to modulate interactions between SR proteins and other splicing factors. Therefore, it is likely that RS domain mediated interactions between SR proteins and factors involved in other RNA metabolic reactions are also subject to regulation via CLK activity. Disruption of CLK phosphorylation of SR protein RS domains may result in a reduction of SR protein-CFIm or SR protein-CstF interaction. Additionally, there is a possibility that disruption of CLK activity will directly reduce phosphorylation of the CFm proteins. Either situation would negatively impact the ability of the 3′ end processing machinery to recognise poly(A) sites and effectively cleave nascent RNA molecules.

1.3 Disruption of RNA Processing in Human Disease

Studies of genetic diseases have often focussed on the protein coding regions of genes, especially mutations changing the amino acid sequence of the translated peptide. Synonymous exonic changes and changes occurring in intronic regions can still lead to gene dysfunction and disease. Up to 50% of mutations contributing to disease affect RNA splicing [28]; 10% directly disrupt splice sites [29].

Essential to splicing is the recognition of splice site signals demarcating intronic sequences. Mutations preventing the identification of splice sites can result in loss of exon recognition [4, 29] and potentially introduce a premature termination codon (PTC), as in the case of familial dysautonomia [30]. MCAD deficiency fatty acid disorder is caused by a mutation that disrupts an ESE in the MCAD gene, resulting in skipping of exon 5 and nonsense-mediated decay of the RNA transcript [31].

Mutations affecting RNA splicing have also been implicated in cancer formation and progression. The splicing factor SF3B1 has been shown in a recent study to be mutated in approximately 20% of patients with myelodysplastic syndromes [32]. In prostate cancer, a mutation creates an ESE in the KLF6 gene and promotes expression of an isoform that accelerates tumour progression [33].

SR proteins have also been associated with cancer. Both SRSF1 and SRSF3 are up-regulated in ovarian and colon cancer, among others [5, 34]. For example, SRSF1 regulates splicing in the oncogene MST1R [35]; Over-expression of SRSF1 increases expression of an MST1R isoform that bestows greater cell motility, which is related to tumour progression.

Similar to RNA splicing, mutations in either poly(A) sites or their *cis*-regulatory sequences can lead to disease [36]. Additionally, misregulation of alternative polyadenylation can cause or exacerbate pathological conditions. For example, cardiac hypertrophy and some cancers are associated with a general preference for the selection of proximal poly(A) sites. It is also possible that SR proteins play a role in disease involving misregulated polyadenylation; they are known to both regulate poly(A) site selection [20] and to be involved in disease.

1.3.1 The splicing machinery as a therapeutic target

The involvement of the RNA splicing machinery in a broad array of diseases makes it a potential target for therapeutic intervention. Two approaches have been identified in the development of therapies for splicing related diseases. One approach uses antisense oligonucleotides to target specific regions of the nascent RNA transcript, thus preventing the expression of pathological RNA and protein isoforms. Another approach uses small molecules to modulate cellular signalling events that regulate splicing.

Antisense oligonucleotides can be designed to complement specific nucleotide sequences within a pre-mRNA. Depending on the sequence targeted, the selection of specific splice junctions or entire exon can be controlled. Isoform expression itself can be adjusted by promoting the degradation of target transcripts, while protein-RNA interactions can be prevented by blocking binding sites, for example, ESEs and ESSs. Antisense oligonucleotides have been successfully used to treat patients with Duchenne's muscular dystrophy [34].

Small molecules can be used to modulate splicing by inhibiting or promoting certain cell signalling pathway events. A well known splicing related signalling event is the post-translational phosphorylation of splicing factors, especially those of the SR protein family. The phosphorylation status of SR proteins affects their ability to promote exon recognition. Inhibitors of proteins known to phosphorylate SR proteins have been recently developed, including KH-CB19 [37] and T3 (unpublished, but used in this project). Both CB19 and T3 target the activity of the CLK family of kinases.

Although there is the potential to use small molecules to treat splicing related diseases, inhibiting components of cellular pathways are likely to have many unintended effects. Aside from potential drug off-targets, splicing regulators (*e.g.* CLKs) are important for the normal splicing of diverse transcript species. To fully comprehend the consequences of small molecule splicing modulation, transcriptomic response must be studied in a systematic manner.

1.4 Detecting and Measuring Changes in the Transcriptome

There are several methods by which cellular RNA can be measured and compared. Recently developed RNA sequencing technologies allow the capture and identification of RNA transcript sequences — including splice junctions — without prior knowledge of their existence or composition. RNA-seq, or "Whole Transcriptome Shotgun Sequencing" samples many short RNA fragments from a population of cells. It uses "next-generation" sequencing technologies to produce reads usually around 30–700 base pairs (bp) in length, depending on the technology used. At the same time, the number of reads produced can be very large — up to hundreds of millions, or even billions of reads per run. The number of bases sequenced allows the quantitative representation of the entire transcriptome.

A common approach to RNA sequencing involves fragmenting the transcriptome, or a subset thereof (*e.g.* only coding, polyadenylated transcripts). The fragments are reverse transcribed to create complimentary DNA (cDNA), which are amplified and then sequenced. During sequencing, either a single end, or both ends of the cDNA can be sequenced. Paired-end sequencing libraries, where both ends of a fragment have been sequenced, have the additional benefit of providing the expected length between each read mate-pair. This information is useful for downstream analysis, including gene and RNA isoform quantification.

Standard RNA-Seq methodologies produce reads with no indication of which DNA strand the RNA fragment was transcribed from. Because there are regions of the genome in which genes on both strands overlap, RNA-Seq reads may not always be unambiguously assigned to one strand or the other. To address this problem, "strand-specific" RNA-seq protocols have been developed [8]. Strand-specific RNA-seq libraries are useful for quantifying transcript expression from genomic regions with genes occurring on both the forward and reverse strands.

A drawback of RNA-Seq methods is the short read length. A single RNA-

Seq read typically cannot unambiguously reveal the structure of the full RNA molecule from which it was produced (Figure 1.4a). This problem is exacerbated by the the presence of multi-exonic genes with multiple alternative isoforms. For example, a read may indicate the skipping of an exon if it maps to the two adjacent exons. However, it may not be useful in identifying alternative splicing decisions made upstream or downstream of that particular exon.

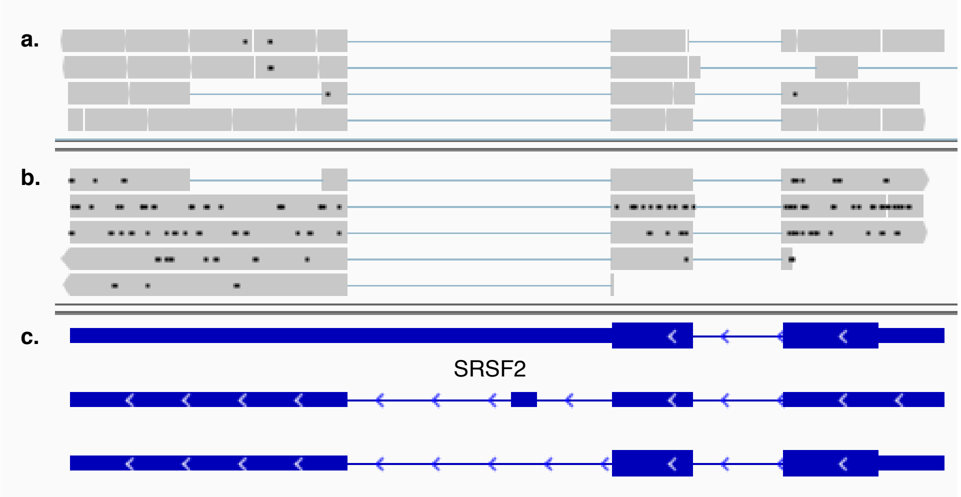

Figure 1.4: A comparison of RNA-Seq vs PacBio cDNA reads mapped to SRSF2 using a plot generated by the Integrative Genomics Viewer. The longer PacBio reads can typically reveal more of a transcript's structure than can single RNA-Seq reads. Grey blocks represent sequencing reads. The thin blue lines between grey blocks represent gaps within reads that are split across introns. Black dots within reads represent deletions. **a**, RNA-Seq reads. **b**, PacBio reads. **c**, SRSF2 transcript structure from RefSeq.

Long read sequencing technologies produce read lengths thousands of base pairs long. The Pacific Biosciences' (PacBio) Single Molecule Real Time (SMRT) technology can produce reads with an average length of 4,200–8,500 bp, with the longest reads reaching greater than 30,000 bp. With these read lengths, large sections of mRNA, or even entire transcripts may be captured (Figure 1.4b).

A potential disadvantage of the PacBio sequencing platform is the error rate: approximately 13% on average for raw reads [38]. However, reads with $\geq 99.9\%$ average accuracy can be constructed from the raw continuous long reads (CLRs): when a single cDNA molecule is sequenced multiple times, the CLRs can be assembled into a single high quality circular conformation sequencing (CCS) read. If a cDNA molecule is too long to be sequenced multiple times before sequencing termination then CLRs representing large portions or even the entire molecule can still be produced, albeit with greatly reduced accuracy.

Another limitation of PacBio sequencing is the moderate throughput. The PacBio RS platform produces around 100 Mb of sequence, while the Illumina HiSeq 2000 can produce 600 Gb [39]. Although short-read sequencing is still preferable for quantitative measurement of transcriptomes, long-read sequencing is valuable for isoform detection and validation.

Current short- and long-read sequencing technologies should be viewed as complementary, rather than as competing, approaches. The high-throughput of RNA-Seq allows the capture of sequence from many distinct RNA species and provides a greater sensitivity than the PacBio platform. RNA-Seq also has a greater per-base accuracy which is critical for mutation detection and accurate identification of splice sites. Therefore, RNA-Seq libraries can be used to predict spliced RNA isoforms with high sensitivity, while PacBio reads can then be used validate the existence of the predicted transcripts.

1.4.1 Computational methods

Extracting information about the transcriptome of a cell population from RNA-Seq libraries is a difficult problem. However, many tools have been developed that attempt to compute statistics from RNA-Seq data, such as gene and RNA isoform expression levels, and relative inclusion levels of alternatively spliced transcript components. Studies using RNA sequencing technologies often follow common analysis workflows starting with read alignment and proceeding to at least one of several different analyses, including differential expression analysis or RNA isoform prediction (see Figure 1.5). Each RNA sequencing method has it's own sources of error and biases that can confound analyses. So, many studies will validate results using an independent approach; For example real-time PCR [40] or Sanger sequencing [41] can be used to verify the existence of spliced isoforms.

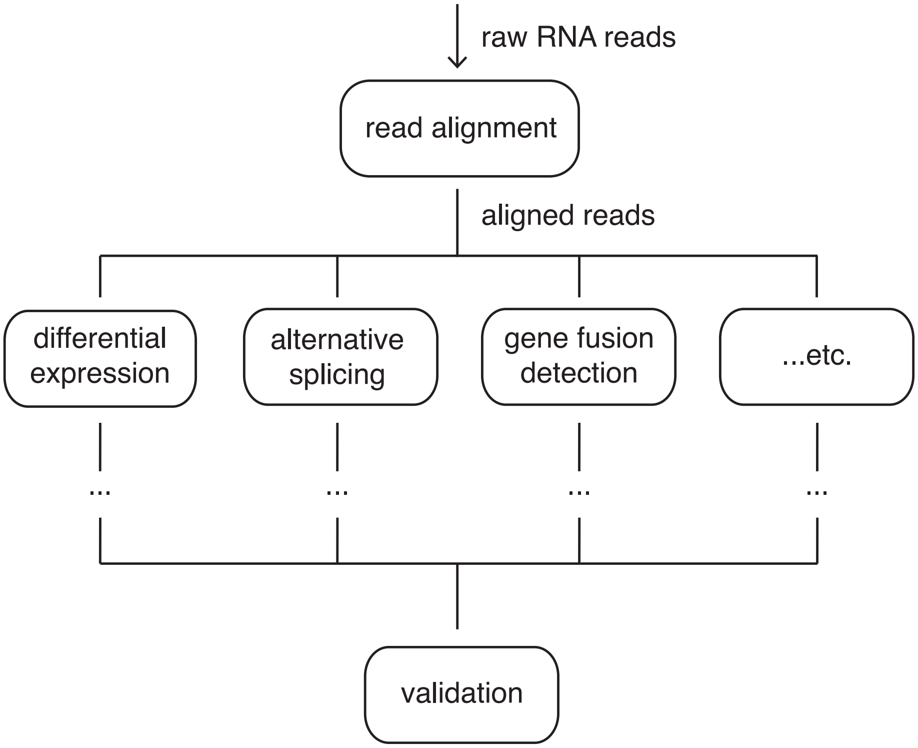

Figure 1.5: Common basic workflow of analysis with RNA sequencing libraries. RNA sequencing reads are first aligned to a reference genome. The resulting aligned reads can then be used in a number of different analyses, including differential expression analysis, alternative splicing quantification, gene fusion detection, etc. The products of these analyses may then be used in further downstream analyses. Validation of results may be performed using a variety of methods.

Splicing-aware RNA sequencing read alignment

The literature describes many methods for accurately aligning DNA sequencing reads to a reference genome. However, determining the genomic origins of RNA reads presents a distinct challenge: RNA reads can represent regions of RNA containing splice junctions. If a read overlapping a splice junction is to be accurately aligned to a reference genome, the read must be split apart and each portion mapped to the corresponding exons. Doing so can be difficult if the split read portions have insufficient sequence specificity to be accurately mapped to the reference genome. DNA sequence aligners are not optimized for the large-gap alignment necessary for RNA read mapping.

A potential solution to the problem of split read alignment is to use reference transcriptome sequences instead of a reference genome. By aligning to a reference transcriptome, the need to split RNA sequencing reads across introns is greatly reduced. However, alignment would be restricted to a set of known or predicted RNA sequences, hindering novel isoform detection. Additionally, reads originating from transcripts not present in the reference transcriptome may be aligned to an incorrect reference transcript.

Rather than align RNA reads to a reference transcriptome, alignment can be performed against both a reference genome and a database of exon junction sequences. This approach eliminates the need for a reference transcriptome and allows the entire genome to be queried for possible matches. But, the set of splice junction sequences is also limited to known or predicted exon junctions, making the alignment of reads containing unknown splice junctions problematic.

These issues motivated the development of methods specifically tailored to RNA sequencing read alignment. Some short-read (*i.e* RNA-Seq) alignment methods, such as GSNAP [42] and STAR [43], are able to detect and map reads across both annotated and predicted splice junctions. However, short-read alignment methods may not be the most appropriate choice for longer reads; For example, the GMAP [44] cDNA aligner is recommended for PacBio reads [45].

Alternative splicing detection and quantification

Common problems in the study of alternative splicing are the identification and quantification of existing spliced isoforms. Methods developed to address these problems employ a variety of techniques to accomplish their objectives, and computations can be performed at the level of individual AS events or at the level of whole alternative transcript isoforms. Some approaches to AS detection and quantification commonly use information inherent in mapped RNA-Seq reads. During aligment to a reference genome, some reads are split and each segment mapped to exonic sequences separated by an intron. These reads are useful for indicating the precise location of exon junctions. When paired-end RNA sequencing data is available, the genomic distance between two mate-pairs mapped to the reference genome can be compared to the expected value of mate-pair distances in the originating sequence library. When mate-pair distances are longer than expected, it is possible that an exon in the gene model has been skipped in the final mRNA molecule. Although mate-pair distances cannot identify the precise location of exon junctions, they are valuable for inferring the exonic architecture of the originating cDNA fragment.

A measure of AS is the percent spliced in (PSI) value

$$PSI = \frac{I}{I+E} \tag{1.1}$$

where I is the number of inclusion isoform transcripts, and E is the number of exclusion isoform transcripts [46]. For example, the inclusion isoform for a SE event would be the isoform containing the potentially skipped exon. PSI values can be compared between two samples to identify RNA isoforms or AS events that are differentially spliced.

A popular method for AS analysis is the MISO software package [46]. MISO calculates PSI values for a set of annotated AS events belonging to 8 different classes (SE, retained intron (RI), MXE, A3SS, A5SS, AFE, ALE, tandem UTR) using a Bayesian approach. When comparing PSI values between two samples, MISO calculates a Bayes factor statistic

$$BF = \frac{\Pr(D|M_1)}{\Pr(D|M_2)} \tag{1.2}$$

where D is the observed data, and M_1, M_2 are two statistical models. The Bayes factor in this application is the likelihood ratio of the observed data being produced under the assumption of differential splicing occurring, over the assumption of no differential splicing. Essentially, the higher the Bayes factor, the more likely it is that differential splicing has occurred. MISO is an appropriate choice for projects requiring differential splicing analysis of a broad range of AS event types in human cells. Although MISO contains only a specific set of functionality, the field of computational AS methods has developed to the point where there exists a number of statistically rigorous tools that can satisfy the needs of most sequencing based AS studies [47].

Gene expression quantification

The simplest way gene expression can be estimated given a RNA-Seq library is to count the number of reads or read pairs mapping to regions of the genome corresponding to annotated gene locations. For some applications, such as differential gene expression analysis using DESeq [48] or edgeR [49], it is necessary to calculate expression using this strategy. However, raw read counts are biased by factors including the sequencing depth of a library, and the length and GC content of genes. Generally, the higher the sequencing depth, or the longer the gene, the more reads will map to that gene. As a result, it is necessary to employ some form of read count normalization when dealing with gene expression analysis.

Some normalization schemes attempt to find a suitable scaling factor used to divide gene read counts within a sequencing library. The DESeq and edgeR packages both use this approach for differential expression analysis. Another approach is to use quantile normalization to transform the gene expression distributions of each RNA-Seq library in such a way as to make them identical. Yet another approach is calculating Reads Per Kilobase of exon model per Million mapped reads (RPKM) values

$$RPKM = \frac{10^9 C}{NL} \tag{1.3}$$

where C is the number of reads mapped to a gene's exons, N is the total number of mapped reads in the sequencing library, and L is the length of the gene's exons in base pairs [50]. RPKM values represent global (rather than relative, *e.g.* PSI) expression level, and normalize read counts by the number of mapped reads in a sequencing library and by the lengths of gene models.

A variant of the RPKM measure, Fragments Per Kilobase of exon model per Million mapped reads (FPKM), is produced by the Cufflinks software [51]. The calculation of FPKM values takes into account that with paired-end sequencing data, only one mate of a read pair originating from the same cDNA fragment might be mapped to the genome reference. This results in the double counting of fragments with both mate-pairs mapped while only counting other fragments once. FPKM attempts to count cDNA fragments rather than individual RNA-Seq reads, thereby reducing this bias. The Cufflinks software can also correct for fragment bias (certain sequences being preferentially selected for by primers during PCR) when calculating FPKM values [52].

1.5 Experimental Approach and Aims

Takeda Pharmaceutical Company Limited has recently developed T3 — a novel compound that suppresses RNA processing by inhibiting CLK phosphorylation of RS domains. The Takeda T3 compound inhibits CLK activity with a greater specificity than previously reported CLK inhibitors [unpublished data]. Although methods for splicing inhibition have been described [37, 53], the transcriptomic effects of progressively disrupting RNA processing have not been assessed in a systematic manner. Using this novel T3 compound, cellular responses to pharmacological restriction of RNA processing can be measured. Concentration-based analysis will facilitate the identification of transcriptomic components sensitive to CLK inhibition, and may provide valuable insight into the importance of RS domain phosphorylation in the RNA processing regulatory landscape.

Alternative splicing can be categorised into eight different event types (Figure 1.3). Each event type may rely on the activity of SR proteins to a greater or lesser extent. SR proteins also have a role in non-splicing reactions, including $3'$-end formation. Additionally, the phosphorylation status of the RS domain-containing CFIm appears to be important to the $3'$-end cleavage reaction. The vulnerability of RNA processing events to CLK inhibition, and the manner in which these events react to progressive repression of CLK is currently unknown.

Individual RNA processing events may have differing responses to T3 treatment. For example, the PSI values of AS events may increase or decrease to varying degrees upon treatment. The direction of response and level of sensitivity may reflect the strength of *cis* regulatory signals, or other relevant RNA sequence characteristics. There have been efforts to characterise an RNA splicing code [54]; Nevertheless, there is a lack of research in transcriptome-wide RNA features predictive of splicing changes caused by the global impedance of SR protein function.

Disruption of RNA processing efficacy can lead to changes in the composition of the transcriptome, which may comprise both changes in RNA isoform balance, as well as gene expression level. These changes may reflect both the direct effects of CLK inhibition, as well as compensatory responses by the cell. For example, disruption of AS may increase production of aberrant transripts, which may then prompt the cell to up-regulate the expression of gene isoforms involved in nonsense-mediated decay. Which biological processes are most vulnerable or responsive to CLK inhibition and alterations in RNA processing efficacy has yet to be described.

RNA processing patterns are dependent on biological context, including cell type [54] and tumour/normal status [5]. CLK phosphorylation of RS domains appears to be fundamental to the process of RNA processing. However there may still be variations between cell types in the degree to which RNA metabolism relies on CLK activity. To gain insight into the regulation of RNA processing via CLK-mediated SR protein phosphorylation, HCT116 and hTERT cells were treated with progressively increasing concentrations of T3. Vehicle-treated cells were used as a negative control. To compare the effects on splicing between T3 treated cells and cells with artificially reduced CLK expression, a CLK small interfering RNA (siRNA) experiment was performed with HCT116 cells. RNA was measured using RNA-Seq and Pacific Biosciences' RS platform (see Section 2.1).

Gene expression and RNA splicing changes were quantified computationally using the MISO [46] and Cufflinks [51] software. Preliminary inspection of treated RNA-Seq libraries revealed the treatment-dependent formation of conjoined transcripts. So, a transcriptome-wide search for conjoined transcripts was performed using a published gene-fusion detection method [55]. Biological processes affected by T3 treatment were found by selecting genes exhibiting changes in splicing or expression to build functional interaction networks [56], which were then queried for enriched GO biological process terms [57]. RNA features associated with splicing changes due to T3 treatment were computed using gene annotations and published sequence motifs [58].

A description of the generated datasets and the results of computational analysis are included in chapter 2 of this document. The results are split up into three main parts. In Section 2.2, the dose depended effects on AS are reported along with the results of an investigation into affected biological processes. This section also compares the changes in AS between the HCT116 and hTERT cell types, as well as between T3 treated and CLK siRNA transfected HCT116 cells. Section 2.3 includes a characterisation of conjoined gene transcripts produced as a result of T3 treatment and biological processes affected by conjoined gene transcription. Finally, Section 2.4 describes the effects of T3-induced CLK inhibition on gene expression and the biological processes affected by differential expression.

Chapter 2

Transcriptomic Consequences of CLK Inhibition

2.1 Datasets

The CLK inhibitor compound, T3, was applied to HCT116 malignant colon epithelial cells and normal hTERT cells at multiple concentrations. RNA was measured using either an unstranded (HCT116 cells) or stranded (both HCT116 and hTERT cells) RNA-Seq protocol, or using Pacific Biosciences SMRT platform [9]. Table 2.1 summarizes the three datasets.

Table 2.1: Summary of T3 treatment datasets. Each dataset contains sequences from either T3 treated or control cell populations. (unstr) and (str) indicates an unstranded or stranded RNA-Seq protocol was used, respectively. An 'X' indicates that a sequencing library exists for the appropriate T3 concentration and dataset.

	HCT116			hTERT
T3 dose (μM)	RNA-Seq (unstr)	RNA-Seq (str)	PacBio	RNA-Seq (str)
0.0	X	X	X	X
0.05	X			
0.10	X			
0.50	X	X	X	X
1.0	X	X		X
5.0	X	X	X	X
10.0	X			

The primary dataset used for analysis was the HCT116 unstranded RNA-Seq dataset. This dataset includes the largest number of T3 treatment observations, providing the ability to detect changes at both very small and large doses as well as providing greater resolution for response pattern detection. The hTERT dataset was used to determine whether observed transcriptomic response patterns were HCT116 cell-type specific, or observable in cells with differing biology. As the hTERT dataset was sequenced using a stranded RNA-Seq protocol, a second HCT116 dataset was generated for comparison using the same RNA-Seq protocol and the same T3 concentrations as the hTERT dataset.

Multiple datasets were also generated with the purpose of validating results presented in this study. A CLK knockdown data set was generated by using siRNA to target each or a combination of the CLK proteins in HCT116 cells. Two control libraries were generated by either knocking down NT3 (a growth factor in neurons), or treating cells with only vehicle Lipofectamine® 2000. The RNA from each CLK siRNA sample was sequenced using RNA-Seq (Table 2.2).

Table 2.2: Summary of CLK siRNA knockdown RNA-Seq libraries. The knockdown experiment was performed using HCT116 cells. An 'X' indicates that the corresponding sequencing library was generated from cells with the indicated target knocked down. (ctrl) indicates a control library, and 'None' represents a sample treated with vehicle Lipofectamine® 2000.

	siRNA target					
Sample	CLK1	CLK2	CLK3	CLK4	NT3 (ctrl)	None (ctrl)
1	X					
2		X				
3			X			
4				X		
5	X	X	X			
6			X	X		
7	X	X		X		
8	X	X	X	X		
9					X	
10						X

Another dataset, consisting of RNA sequences obtained from the PacBio RS platform, was generated mainly for the purposes of validating the existence of spliced isoforms arising due to CLK inhibition. The PacBio sequencing platform is suited for RNA isoform detection as it is able to produce long reads, enabling the identification of large portions of transcript structure. The Pacbio dataset includes both high-quality CCS reads, and lower quality CLR sequences.

All of the RNA-Seq libraries were aligned using GSNAP. The aligned libraries were then processed to remove potential PCR duplicates. The PacBio libraries were aligned with the GMAP aligner and filtered to only include reads whose aligned proportion is at least 90%, and have at least 80% identity with the reference.

2.2 T3 Treatment Induces Dose-Dependent Alternative Splicing Changes

Relative inclusion levels of alternative splicing events were quantified using MISO in all three T3-treated RNA-Seq datasets. The resulting PSI values in each treated library were compared to corresponding control PSI values. Alternative splicing events were called as differentially spliced if the MISO-calculated Bayes factor (Equation 1.2) was ≥ 20 and the difference in PSI values between treated and untreated samples was ≥ 0.1.

To assess the transcriptome's sensitivity to CLK inhibition, the number of differentially spliced events were counted for each treated library. The number of differentially spliced events increased with higher dosage (Figure 2.1). This response pattern demonstrates that the Takeda T3 compound is able to inhibit the splicing of a large number of exon junctions. A large change in the number of affected events occurred at 0.50μM (4474 events compared with 799 and 1088 for 0.05μM and 0.10μM, respectively, or 5.6 and 4.1 fold more events), suggesting that at this concentration a regulatory mechanism was disrupted, resulting in greater numbers of differentially spliced transcripts.

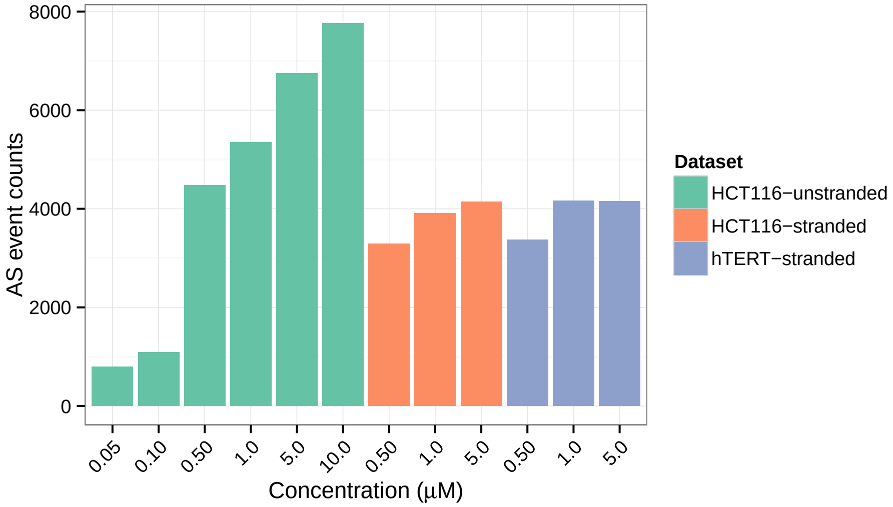

Figure 2.1: Differentially spliced event counts for the HCT116 and hTERT datasets. Events have a Bayes-factor >= 20

2.2.1 Alternative splicing response to CLK inhibition is common to both HCT116 and hTERT cell types

Splicing response to CLK inhibition between HCT116 and non-malignant hTERT cells was compared to ascertain the extent to which biological context affects the reliance of splicing on normal CLK activity. Treated hTERT and HCT116 cell transcriptomes were sequenced with a stranded RNA-Seq protocol (see Table 2.1). To investigate the degree of overlap between differentially spliced AS events in the three T3-treated RNA-Seq datasets, the events affected by T3 treatment were collected for each dataset. The number of overlapping and dataset-specific events were then counted (Figure 2.2).

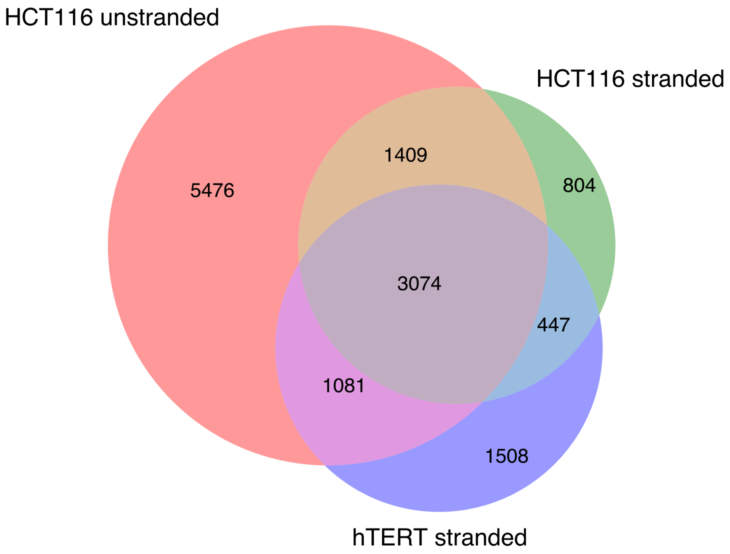

Figure 2.2: Venn diagram illustrating the number of unique overlapping and dataset-specific differentially spliced MISO events between the HCT116 and hTERT RNA-Seq datasets.

The HCT116 unstranded RNA-Seq dataset produced the greatest number of differentially spliced AS events (11,040), followed by the hTERT (6,110) and HCT116 stranded RNA-Seq datasets (5,734) (Figure 2.2). However, the unstranded RNA-Seq dataset includes more treated samples, including the 10.0µM concentration. The large majority of events for both stranded RNA-Seq datasets overlap with the events from at least one other dataset (HCT116: 86%; hTERT: 75%), while only 50% of events from the HCT116 unstranded RNA-Seq dataset overlap with the events from another dataset (Figure 2.2). The stranded RNA-Seq libraries may include less splicing information than the unstranded RNA-Seq libraries (see Section 2.5). Of the differentially spliced events detected in hTERT cells, 75% were also detected in HCT116 cells. 37% of all HCT116 events and 61% of events from the HCT116 stranded RNA-Seq dataset were also detected in hTERT cells. The amount of AS event overlap between hTERT and HCT116 cells suggests that the effects of CLK inhibition are not predominantly hTERT or HCT116 cell-type specific.

2.2.2 Splicing and cell cycle related genes are sensitive to CLK inhibition

Identifying genes differentially spliced at low T3 concentrations will point towards the biological processes most sensitive to loss of CLK activity. Additionally, it may hint at novel roles for CLK phosphorylation in non-splicing processes. Observing affects only occurring at higher concentrations may reveal how the cell responds to widespread RNA processing disruption.

Affected biological processes were determined by identifying differentially spliced genes for each T3 concentration. Genes were then grouped according to whether they were differentially spliced in the 0.05−0.5μM or 1.0−10.0μM CLK inhibitor treated samples. Each group of genes was used to create a gene interaction network using the ReactomeFI Cytoscape plugin [56]. Gene interaction networks were queried for enriched GO biological process terms with false discovery rate controlled at 0.05. Each group of significantly enriched biological process gene sets was then used to generate an enrichment map [59] (Figure 2.3, Figure 2.4, Figure 2.5).

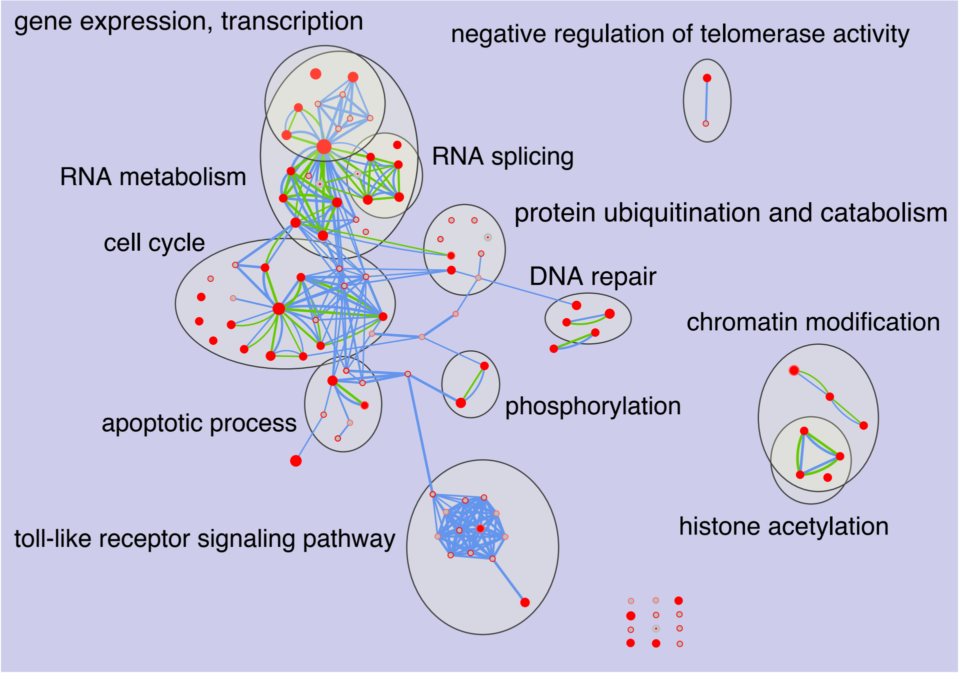
gene expression, transcription
negative regulation of telomerase activity
RNA splicing
RNA metabolism
protein ubiquitination and catabolism
cell cycle
DNA repair
chromatin modification
phosphorylation
apoptotic process
histone acetylation
toll-like receptor signaling pathway

Figure 2.3: Biological process enrichment map for differentially spliced genes in the HCT116 unstranded RNA-Seq dataset. Each node represents a GO biological process gene set. Node cores are coloured red when that gene set is enriched among genes differentially spliced in the the 0.05–0.5μM samples, and the outer ring is coloured red when that gene set is enriched in the 1.0–10.0μM samples. Edge thickness indicates the level of overlap between two gene sets, considering the set of differentially spliced genes in the 0.05–0.5μM (green edges) or 1.0–10.0μM (blue edges) samples.

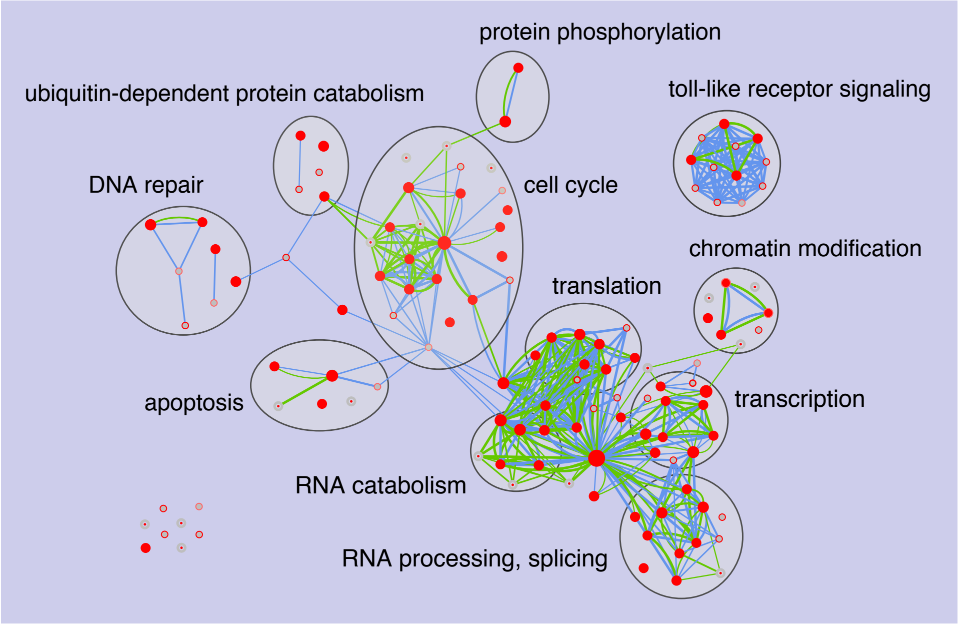

Figure 2.4: Biological process enrichment map for differentially spliced genes in the HCT116 stranded dataset. Each node represents a GO biological process gene set. Node cores are coloured red when that gene set is enriched among genes differentially spliced in the the 0.05μM samples, and the outer ring is coloured red when enriched in the 1.0−5.0μM samples. Edge thickness indicates the level of overlap between two gene sets, considering the set of differentially spliced genes in the 0.05μM (green edges) or 1.0−5.0μM (blue edges) samples.

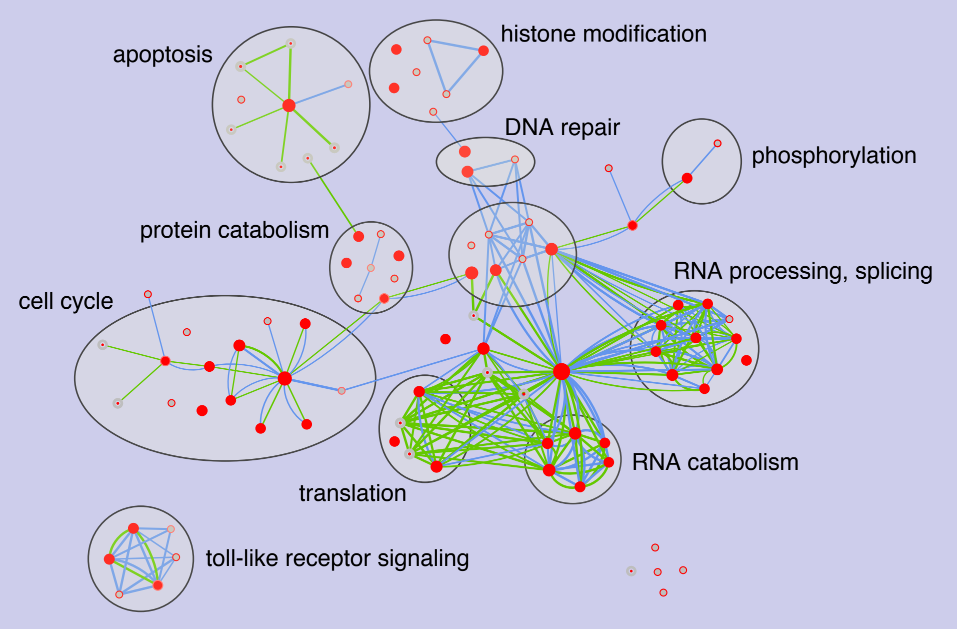

Figure 2.5: Biological process enrichment map for differentially spliced genes in the hTERT dataset. Each node represents a GO biological process gene set. Node cores are coloured red when that gene set is enriched among genes differentially spliced in the the 0.05μM samples, and the outer ring is coloured red when enriched in the 1.0–5.0μM samples. Edge thickness indicates the level of overlap between two gene sets, considering the set of differentially spliced genes in the 0.05μM (green edges) or 1.0–5.0μM (blue edges) samples.

Splicing factors were found to be affected by differential splicing in the lower T3 concentration samples. While the splicing machinery is known to be subject to autoregulation [60], that splicing factors are among the genes affected by even low doses of CLK inhibitor indicates that splicing autoregulatory processes are sensitive to changes in CLK activity. Other forms of RNA metabolism were also affected at lower T3 concentrations, including gene expression and transcription. Cell cycle related genes were also found to be sensitive to CLK inhibition; cell cycle progression is known to rely on the normal operation of RNA splicing [61, 62]. Some groups of related biological processes (*e.g.* those involved with transcription or the cell cycle) had gene sets that were affected at only the higher T3 concentrations. This may be the result of a progressively stronger disruption of these biological processes with increasing T3 dose. A group of genes involved in toll-like receptor signaling were found to be predominantly affected in the 1.0–10.0μM samples. This effect may be an innate immune response to toll-like receptor ligands released from cells dying [63] due to high concentrations of CLK inhibitor. Genes involved in apoptosis are differentially spliced due to treatment, which may also indicate cellular lethality at higher T3 concentrations.

2.2.3 CLK knockdown partially reproduces effects of T3 treatment

The T3 compound prevents CLKs from phosphorylating their target RNA processing factors. Therefore, one may hypothesize that reducing the expression of CLK genes would have a similar effect on RNA splicing. To test this notion, CLK expression was knocked down via siRNA in HCT116 cells and the resulting transcriptomes sequenced using RNA-Seq (Table 2.2).

The RNA-Seq libraries from the CLK knockdown experiment were analyzed with MISO; Each CLK knockdown library and the vehicle control library were compared to the NT3 siRNA control. Differentially spliced AS events were called at a Bayes factor (Equation 1.2) threshold of 20, and PSI change threshold of 0.1, similar to the T3 concentration curve experiment (Section 2.2). A list was compiled of MISO events found to be differentially spliced in any of the CLK siRNA libraries but not in the vehicle control library. This list was then compared to lists of differentially spliced events from the T3-treated HCT116 datasets (Figure 2.6).

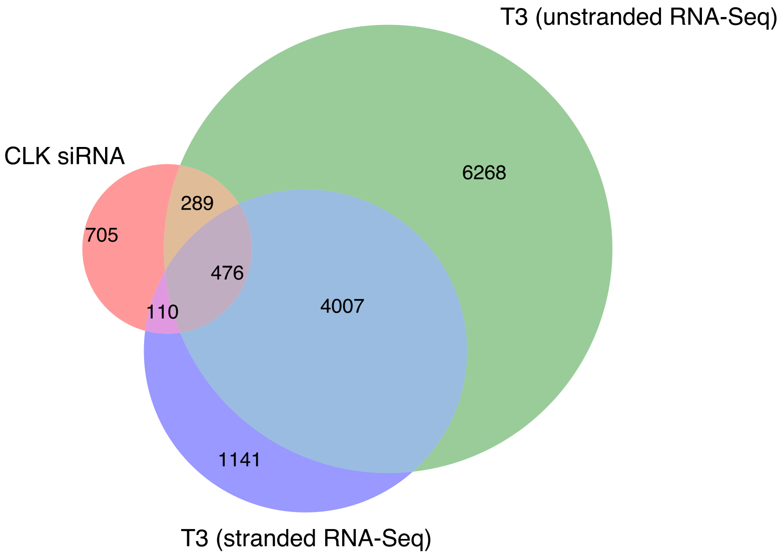

Figure 2.6: Venn diagram showing the number of dataset-specific and common AS events for the CLK knockdown and T3-treated HCT116 datasets.

In total, 1580 unique AS events were found to be differentially spliced in any of the CLK knockdown libraries. Of these events, 875 (55%) were found in at least one of the two T3-treated HCT116 AS event lists, demonstrating that at least some of the effects of T3 treatment are due to loss of CLK function as opposed to inhibition of other targets. Almost half of the events resulting from CLK knockdown were not found to be differentially spliced in the T3 treated datasets. This observation can be partially explained by differences in biological response to depleting CLK RNA versus inhibiting CLK phosphorylation activity.

Genes differentially spliced in both T3 treated cells and cells transfected with CLK siRNA are likely to be specifically affected by loss of CLK activity. Biological processes likely to be affected by splicing changes in this common set of genes were identified by constructing a gene interaction network with the ReactomeFI Cytoscape plugin [56]. Functional enrichment analysis was then performed using the genes in the network (Table 2.3). Biological processes enriched among genes differentially spliced in both T3 treated and CLK siRNA transfected cells included “gene expression”, “mitotic cell cycle”, “chromatin modification”, and “nuclear mRNA splicing, via spliceosome”. The enrichment of these biological processes underscores their sensitivity to normal CLK activity.

Table 2.3: Enrichment of GO biological process terms in differentially spliced genes common between T3 treated and CLK siRNA transfected HCT116 cells.

Biological Process	FDR	Genes
gene expression	0.001	XPO1, THRA, RPL13, U2AF1, RPL10, PTBP1, RPS18, MED15, SRSF11, HSPA1A, HNRNPL, UBE2D3, EIF3B, HNRNPK, TEAD4, RPL10A, RPS24, EEF1A1, CSTF3, EIF4H, NCOR1, NCOR2, RPS2, SNAPC5, POLR1C, EIF4A2, EEF1D, SNRNP70, GTF3C2
mitotic cell cycle	0.0165	XPO1, CEP78, CDC16, NDEL1, CNTRL, AZI1, TFDP1, CDC23, POLD2, AKAP9, POM121, PPP1R12A, ODF2, BUB3, LMNA, CEP63, CSNK1E
chromatin modification	0.021	MORF4L2, MTF2, HDAC5, NCOR1, MBTD1, CHD9, CHD3, PHF19
translational initiation	0.024	RPL13, RPL10, RPS18, EIF3B, RPL10A, RPS24, EIF4H, RPS2, EIF4A2
translational elongation	0.02525	RPL13, RPL10, RPS18, RPL10A, RPS24, EEF1A1, RPS2, EEF1D
nuclear mRNA splicing, via spliceosome	0.03183	U2AF1, PTBP1, SRSF10, SRSF11, HNRNPL, HNRNPK, CSTF3, DDX5, SF1, SNRNP70

2.2.4 T3 induced CLK inhibition reduces splice junction recognition efficacy

CLK inhibition causes changes in splicing levels in many alternatively spliced exons. In addition, multiple AS event types exhibited changes in splicing patterns upon T3 treatment. Understanding the manner in which each AS event type responds to CLK inhibition may provide insight into regulatory differences between event types. Specifically, relative sensitivities to splicing factor phosphorylation status may be revealed.

Differentially spliced events were identified and their PSI values in each T3 concentration were collected. Events with missing PSI values were removed. By inspecting the PSI value distributions at each T3 concentration, several response patterns can be observed (Figure 2.7, Figure 2.8, Figure 2.9). First, the PSI values of SE events decrease as drug concentration is increased (medians: -0.02, -0.13, -0.18, for 0.10μM, 0.50μM, and 1.0μM), indicating that these exons are being skipped more often due to treatment. The most substantial PSI decrease occurs at the 0.50μM concentration (6.5 fold decrease in median PSI from 0.10μM). This observation supports the notion that the 0.50μM concentration surpasses a biological threshold, resulting in widespread structural changes within the transcriptome. RI events tend to increase in PSI over increasing CLK inhibitor concentration (medians: 0.0, 0.02, 0.09, for 0.10μM, 0.50μM, and 1.0μM), demonstrating a tendency for introns to be retained more often as a result of treatment. However, retained introns see a more substantial increase in PSI at 1.0μM, compared to 0.50μM (4.5 fold increase in median PSI from 0.50μM). This response pattern suggests that intron retention is more resilient to CLK inhibition compared to exon skipping. In contrast to SE and RI events, A3SS and A5SS events both see a more gradual increase in PSI. Increases in alternative splice site PSI represents a tendency towards including an exon's extension (*i.e.* choosing a splice site farther away from the centre of the exon).

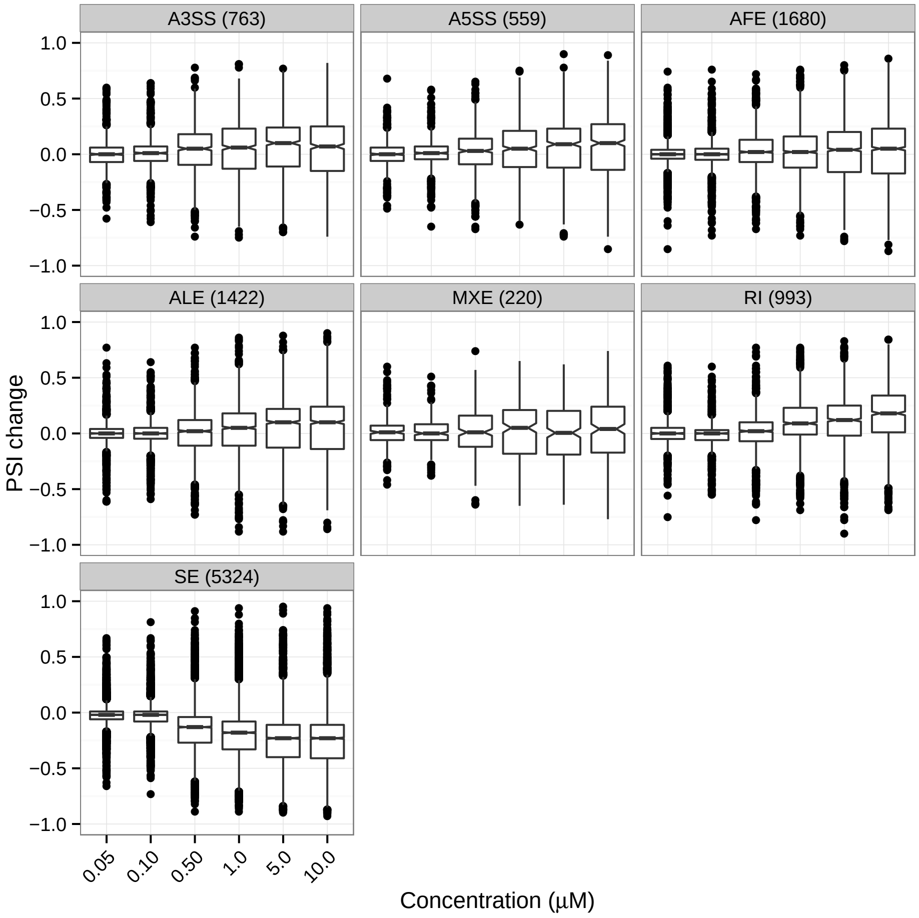

Figure 2.7: AS event type PSI distributions across CLK inhibitor concentration for the HCT116 unstranded RNA-Seq dataset. The number of events for each event type is shown in parentheses. Notches extend $\pm 1.58 \frac{IQR}{\sqrt{n}}$, where IQR is the inter-quartile range.

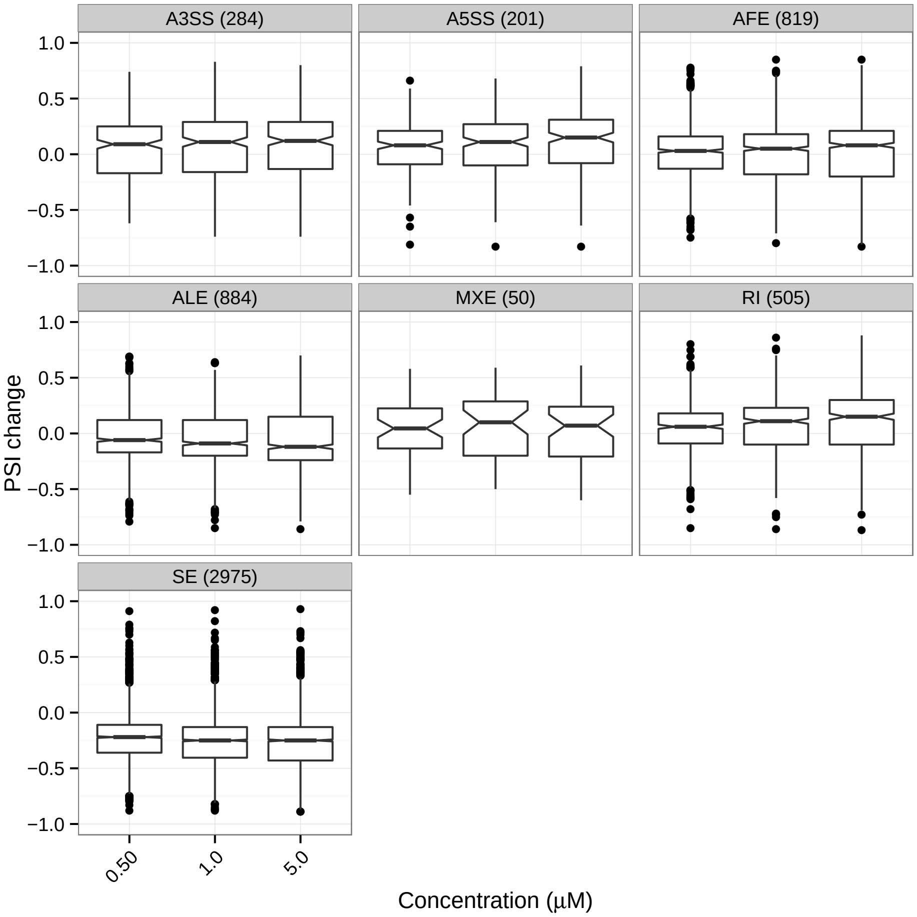

Figure 2.8: AS event type PSI distributions across CLK inhibitor concentration for the HCT116 stranded RNA-Seq dataset. The number of events for each event type is shown in parentheses. Notches extend $\pm 1.58 \frac{IQR}{\sqrt{n}}$, where IQR is the inter-quartile range.

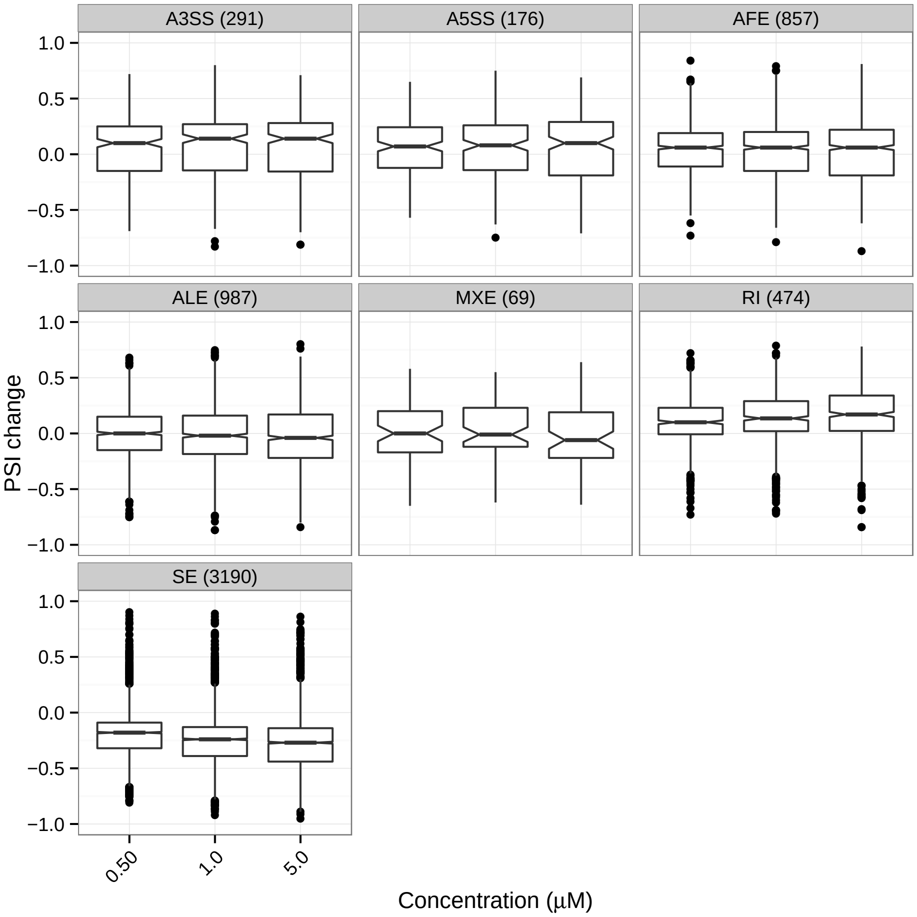

Figure 2.9: AS event type PSI distributions across CLK inhibitor concentration for the hTERT stranded RNA-Seq dataset. The number of events for each event type is shown in parentheses. Notches extend $\pm 1.58 \frac{IQR}{\sqrt{n}}$, where IQR is the inter-quartile range.

The stranded RNA-Seq datasets were unable to show a shift in SE PSI change distributions at 0.5µM due to the lack of samples that have been treated with T3 concentrations lower than 0.5µM. The dose-dependent changes in RI (and other event type) PSI distributions were often not as apparent in the stranded RNA-Seq datasets, partly due to the lack of treatment points, but possibly also due to reduced splicing information in the stranded RNA-Seq libraries compared to the unstranded RNA-Seq libraries (see Section 2.5).

The variety in response patterns reveal that different AS event types have varying levels of sensitivity to CLK inhibition. From these observations, it may be concluded that different classes of alternative splicing events are regulated through different mechanisms that in turn exhibit varying levels of sensitivity to CLK phosphorylation efficacy. Further, the tendency of the splicing machinery to select the exclusion isoform of SE events and the inclusion isoform of RI events suggests that CLK inhibition reduces splice site recognition.

2.2.5 PSI clustering reveals distinct AS response groups

In aggregate, AS events respond to CLK inhibition following event type determined patterns. However, enforcing an event type based segregation of AS events may be concealing finer-grained response profiles. Clustering of AS event PSI profiles will reveal treatment response patterns in an event class unaware manner.

Clustering of PSI profiles was performed using the WGCNA [64] clustering tool. Events were selected for clustering if they were differentially spliced in any of the treated samples at a Bayes factor threshold of 20. Events with missing PSI values were removed unless they contained only two non-consecutive missing values in the case of the HCT116 unstranded RNA-Seq dataset, or only one missing value in the stranded RNA-Seq datasets. Missing values were replaced using linear interpolation. WGCNA was run with networkType="signed" and minModuleSize=25. The WGCNA clustering package requires a soft threshold value which can be chosen by attempting to maximise both the scale independence and connectivity of the PSI correlation network. Soft thresholds of 17, 28, and 24 were selected for the HCT116 unstranded and stranded RNA-Seq, and hTERT datasets, respectively (Figure A.1, Figure A.2, Figure A.3). The threshold for the HCT116 unstranded RNA-Seq dataset was chosen by selecting one of the thresholds where the scale free topology model fit starts to plateau on the model fit versus threshold curve. For the stranded RNA-Seq datasets, values above 20 were chosen as this produced visually distinct clusters and agrees with the suggested guidelines for threshold selection when model fit R^2 values do not reach above 0.8 [65]. The scale free topology model fit can be low when clustering time-series data [65], which the CLK inhibitor concentration curve data can be considered to be. A representative event (*i.e.* "eigenevent") was calculated for each cluster, and events whose PSI profiles did not strongly correlate with the eigenevent (Pearson correlation coefficient ≥ 0.75) were removed.

Clustering revealed several distinct response patterns common across multiple event types and both cell types (Figure 2.10, Figure 2.11, Figure 2.12, number of events per cluster shown in plots). This resulted in 28 distinct PSI profile clusters for the HCT116 unstranded RNA-Seq dataset, and 7 clusters for the two stranded RNA-Seq datasets. Similarities in clustered PSI response patterns can be observed between the HCT116 and hTERT cell types when considering the two stranded RNA-Seq datasets. Similar response patterns can also be observed in the HCT116 unstranded RNA-Seq dataset, although the PSI response patterns in this dataset will be somewhat different due to differences in the number of observations. A summary of proposed similar clusters between the three datasets is included in Table 2.4. Common response patterns found in the three RNA-Seq datasets are likely genuine.

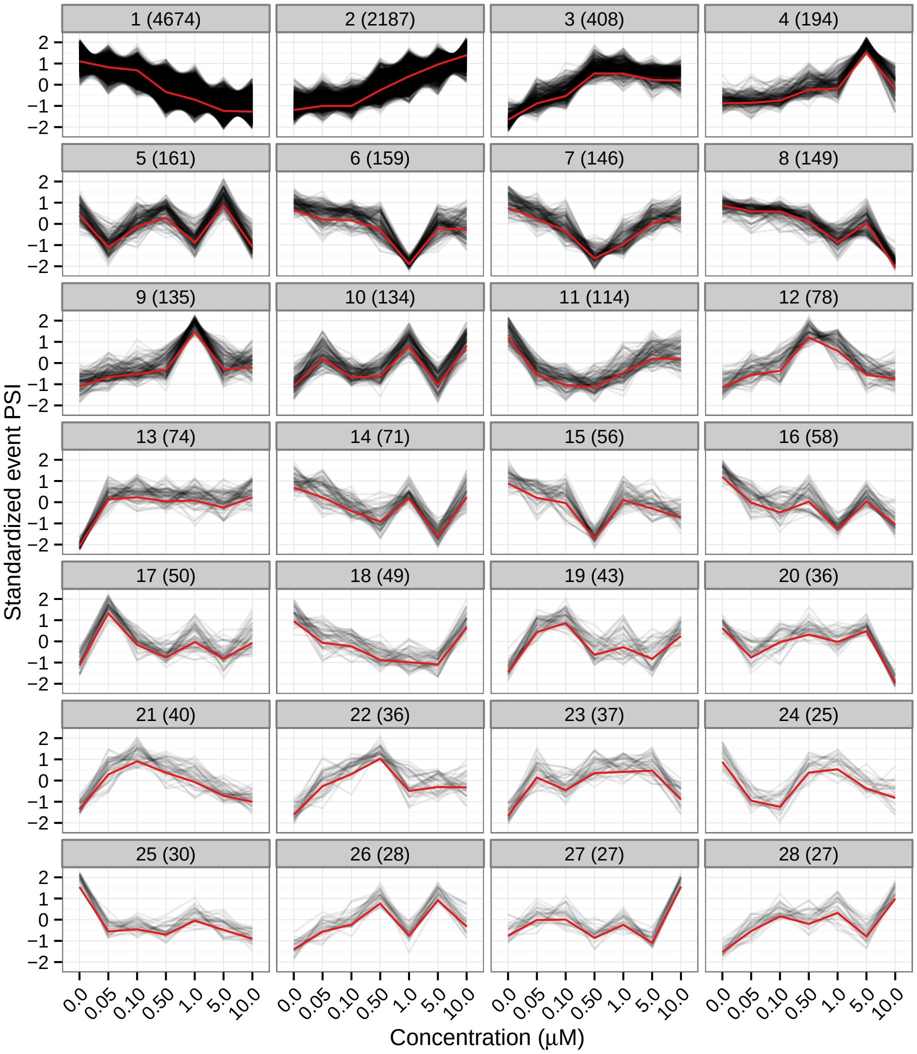

Figure 2.10: AS event PSI clusters for the HCT116 unstranded RNA-Seq dataset. Black lines represent AS event PSI profiles. Red lines are cluster eigen-events. The number of events in each cluster is shown in parentheses in the cluster label.

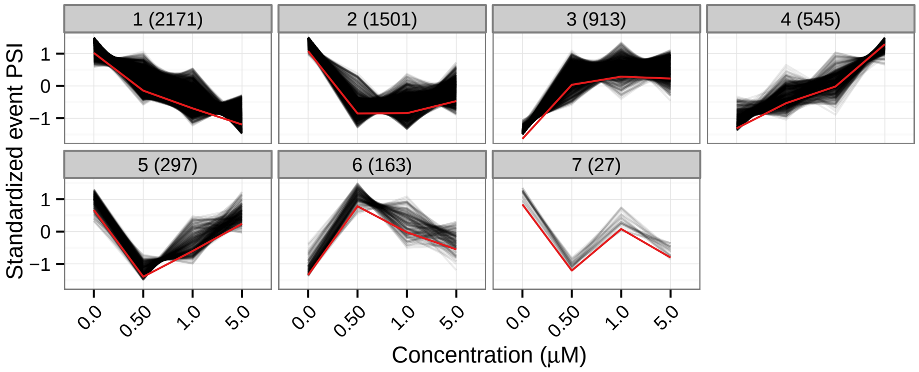

Figure 2.11: AS event PSI clusters for the HCT116 stranded RNA-Seq dataset. Black lines represent AS event PSI profiles. Red lines are cluster eigen-events. The number of events in each cluster is shown in parentheses in the cluster label.

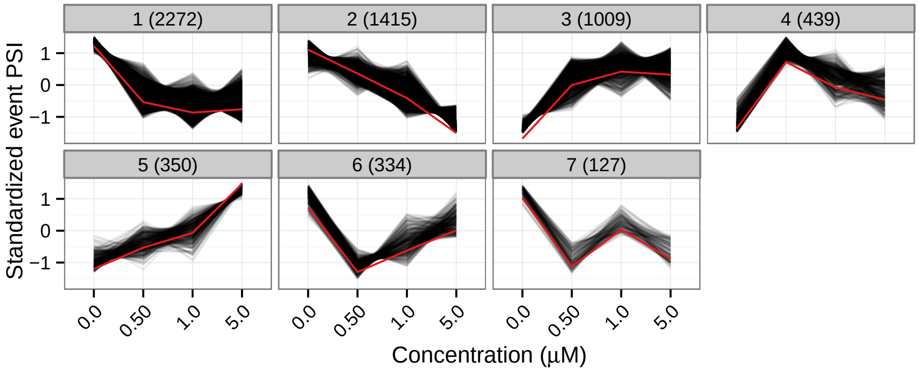

Figure 2.12: AS event PSI clusters for the hTERT stranded RNA-Seq dataset. Black lines represent AS event PSI profiles. Red lines are cluster eigen-events. The number of events in each cluster is shown in parentheses in the cluster label.

Table 2.4: Proposed similar AS PSI response clusters between the three RNA-Seq datasets. (unstr) and (str) indicates an unstranded or stranded RNA-Seq protocol was used, respectively.

HCT116		hTERT
RNA-Seq (unstr)	RNA-Seq (str)	RNA-Seq (str)
1	1	2
1	2	1
3	3	3
2	4	5
7	5	6
12	6	4
25	7	7

As PSI clustering was performed in an AS event type unaware manner, each cluster may contain a variety of event types. Calculating cluster event type proportions revealed a variety of event type distributions between clusters (Figure 2.13). Each cluster was enriched for certain event types, compared with the distribution of all differentially spliced events chosen for clustering (Table 2.5, Table 2.6, Table 2.7). General distributional trends are most apparent when inspecting the event type distributions of the two stranded RNA-Seq datasets. Clusters enriched for SE events (1, 2, 5, and 7 for HCT116; 1, 2, and 6 for hTERT) are characterized by a decrease in PSI between untreated samples and samples treated with 0.50μM of T3 CLK inhibitor. After the 0.50μM concentration, these clusters may either increase or decrease in PSI. The remaining clusters (excluding hTERT cluster 7) are characterised by an increase in PSI between untreated samples and samples treated with 0.50μM of T3. These clusters have a lower proportion of SE events and are enriched for ALE, AFE, A5SS, A3SS, MXE, and RI events. These results agree with the previous observation that SE events tend to decrease in PSI in a T3 dose-dependent manner, while RI events tend to increase in PSI with T3 treatment. A likely cause of this pattern is loss of splice junction recognition efficacy in T3 treated cells.

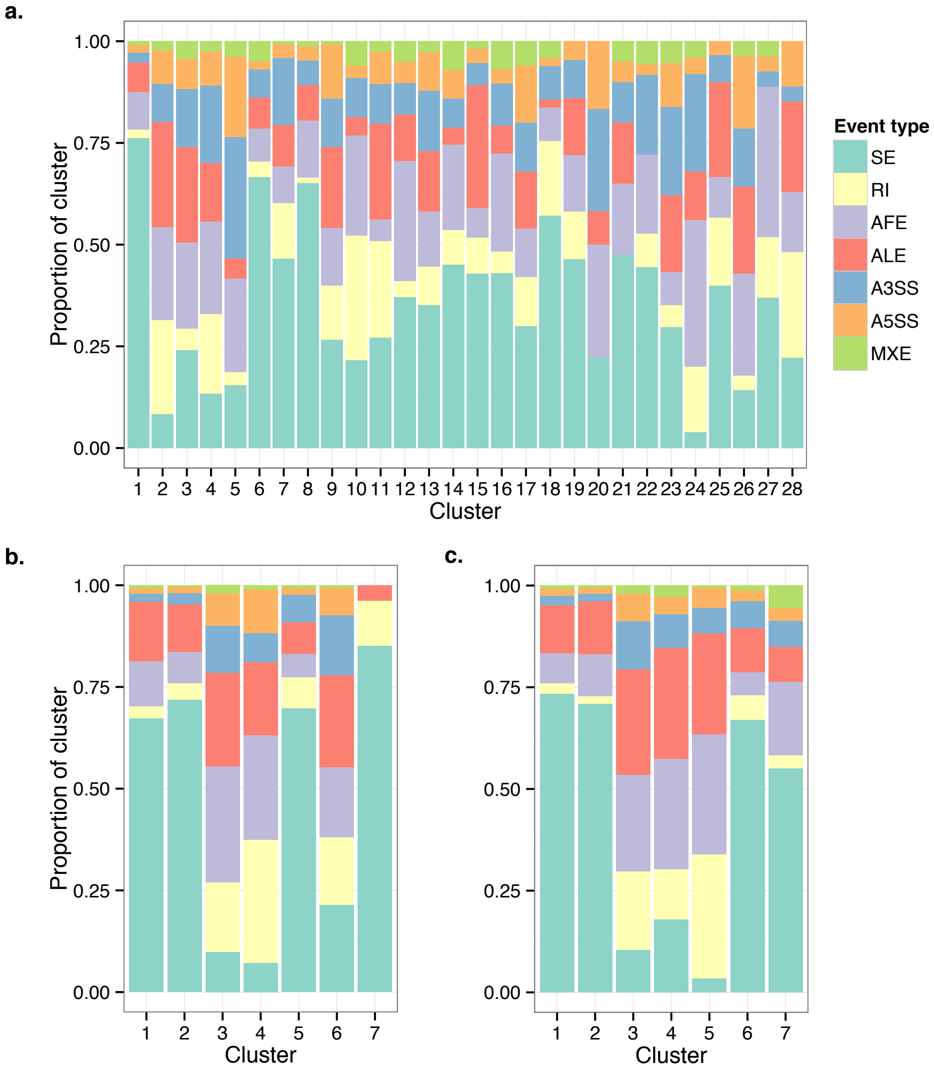

Figure 2.13: AS event type proportions across AS PSI clusters. **a**, HCT116 unstranded RNA-Seq dataset. **b**, HCT116 stranded RNA-Seq dataset. **c**, hTERT dataset.

Table 2.5: AS PSI cluster event type proportion enrichment for the HCT116 unstranded RNA-Seq datasets. Benjamini-Hochberg adjusted p-values from hypergeometric tests are shown if they are below 0.05.

Cluster	A3SS	A5SS	AFE	ALE	MXE	RI	SE
1							0.0
2	$1.98e^{-06}$	$1.97e^{-14}$	$5.06e^{-32}$	$5.05e^{-75}$		$1.71e^{-123}$	
3	$4.97e^{-07}$	0.0364	0.00107	$7.25e^{-08}$	0.00203		
4	$6.11e^{-08}$		0.00582			$3.05e^{-05}$	
5	$3.16e^{-18}$	$1.99e^{-11}$	0.00871				
6					0.0155		$3.64e^{-05}$
7	0.000222						
8							0.000339
9	0.0477	0.000271		0.0477			
10			0.00530		0.00638	$1.06e^{-11}$	
11				0.00582		$1.54e^{-05}$	
12			0.00181				
13	0.0211						
14					0.0120		
15				0.00201			
16					0.0219		
17		0.0130					
18							
19							
20	0.00107	0.00871	0.0477				
21							
22	0.0133						
23	0.00553						
24	0.00740		0.00997				
25							
26		0.0106					

Continued

Cluster	A3SS	A5SS	AFE	ALE	MXE	RI	SE
27			0.00582				
28						0.0125	

Table 2.6: AS PSI cluster event type proportion enrichment for the HCT116 stranded RNA-Seq datasets. Benjamini-Hochberg adjusted p-values from hypergeometric tests are shown if they are below 0.05.

Cluster	A3SS	A5SS	AFE	ALE	MXE	RI	SE
1							$6.63e^{-72}$
2							$6.04e^{-72}$
3	$1.75e^{-19}$	$1.49e^{-12}$	$3.21e^{-35}$	$3.96e^{-11}$	0.000305	$5.30e^{-18}$	
4	0.0128	$3.88e^{-15}$	$2.09e^{-13}$			$5.66e^{-53}$	
5							$4.72e^{-10}$
6	$8.32e^{-07}$	0.0340		0.0128		0.00153	
7							0.000224

Table 2.7: AS PSI cluster event type proportion enrichment for the hTERT. Benjamini-Hochberg adjusted p-values from hypergeometric tests are shown if they are below 0.05.

cluster	A3SS	A5SS	AFE	ALE	MXE	RI	SE
1							7.88e-136
2							8.42e-54
3	8.76e-23	1.71e-12	3.64e-21	5.70e-18	0.0130	2.20e-39	
4	0.00170		2.23e-14	1.32e-09	0.00454	0.000766	
5		0.0407	5.00e-15	2.70e-05		1.00e-38	
6							1.75e-07
7					0.000378		

Analysis of PSI change patterns revealed that AFE and possibly ALE events tend to increase with T3 treatment, although PSI did not always clearly change as a function of T3 dose (Section 2.2.4). However, the event type unaware PSI profile clustering shows that both AFE and ALE events are more prevalent in clusters that increase in PSI in a dose-dependent manner. AFE and ALE events are also present in clusters of events that decrease in PSI, which might explain the lack of a clear dose response in the event type PSI change analysis (Section 2.2.4). PSI increases for AFE and ALE events indicate that an isoform beginning closer to the gene centre is being chosen more often.

2.2.6 ESE density is predictive of splicing response to CLK inhibition

AS events of a particular type with contrasting responses (i.e. increasing vs. decreasing with treatment) may contain differences in splicing signals within overlapping and nearby RNA sequences. One common class of splicing signal is the exonic splicing enhancer (ESE). ESEs are recognized by SR proteins, usually to promote recruitment of the spliceosome to exon junctions. The ESE sequence motifs are degenerate and common in exonic sequences, especially near exon junctions [58]. Exons with a higher density of ESE motifs would present more opportunities for SR proteins to bind to the RNA substrate, promoting inclusion of the corresponding exon.

To test whether ESE density can explain some difference in splicing response, SE and RI events from clusters 1 (SE: 4208 events, RI: 145 events) and 2 (SE: 241 events, RI: 584 events) from the HCT116 unstranded RNA-Seq dataset AS event clustering (Figure 2.10) were selected. These events increase (cluster 2) or decrease (cluster 1) in PSI with T3 treatment. Alternatively included regions in each group of events were queried for the presence of SRSF1, SRSF2, SRSF5, and SRSF6 binding motifs obtained from ESEfinder [58]. The ESE motif search was performed in a probabalistic manner, correcting for background nucleotide rates.

The density of each binding motif was calculated for each sequence. SRSF1, SRSF2, and SRSF5 motif density was significantly higher (one-tailed t-test, Table 2.8) in PSI-increasing vs. PSI-decreasing skipped exons (Figure 2.14). For retained introns, SRSF1, SRSF2, and SRSF6 binding motif density was significantly higher (one-tailed t-test, Table 2.8) in PSI-increasing events (Figure 2.15).

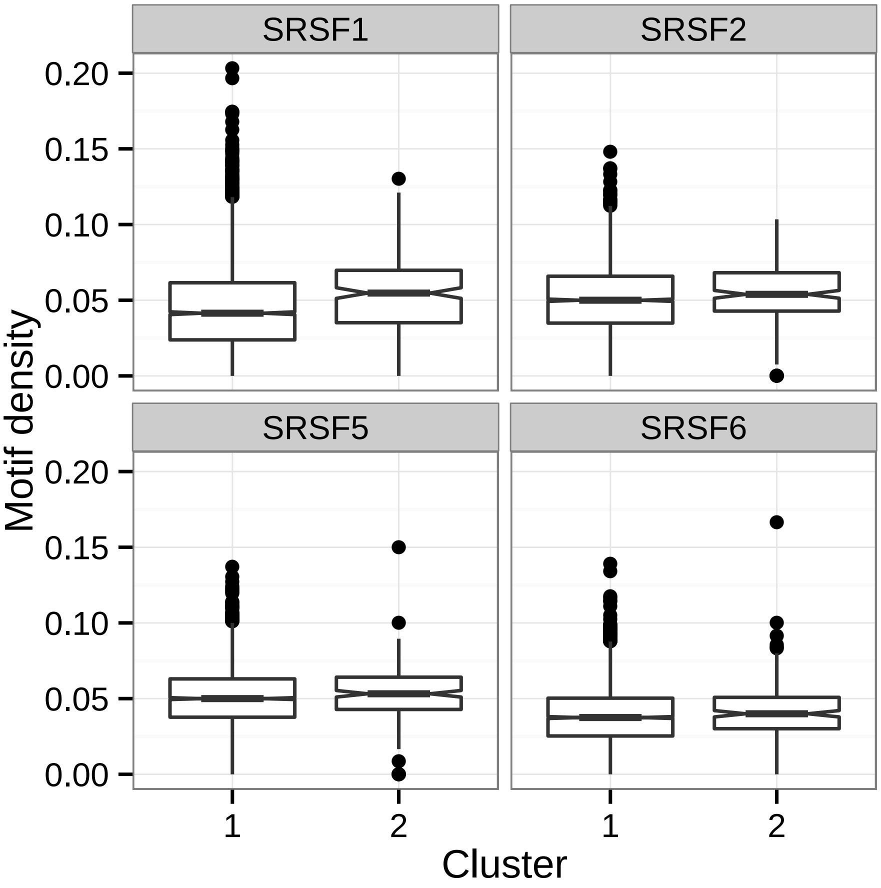

Figure 2.14: ESE density boxplots for skipped exons increasing (cluster 2) or decreasing (cluster 1) in PSI with T3 treatment. ESEs tested include binding motifs for SRSF1, SRSF2, SRSF5, and SRSF6. Cluster 1: 4208 events, cluster 2: 241 events. Notches extend $\pm 1.58\frac{IQR}{\sqrt{n}}$, where IQR is the inter-quartile range.

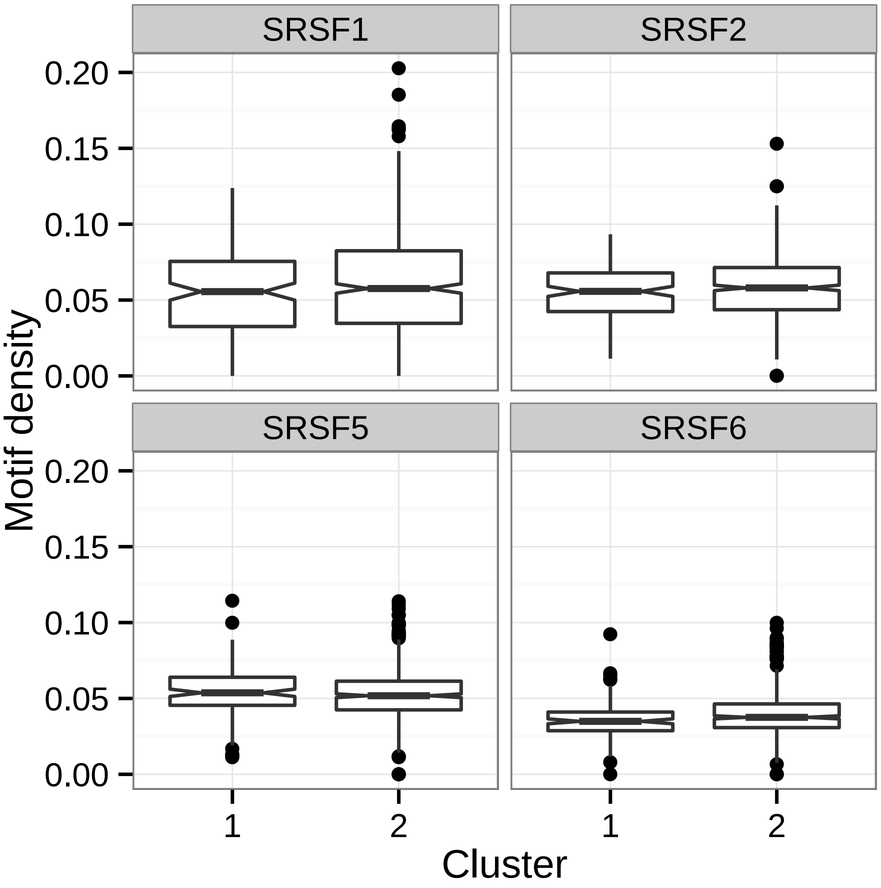

Figure 2.15: ESE density boxplots for retained introns increasing (cluster 2) or decreasing (cluster 1) in PSI with T3 treatment. ESEs tested include binding motifs for SRSF1, SRSF2, SRSF5, and SRSF6. Cluster 1: 145 events, cluster 2: 584 events. Notches extend $\pm 1.58 \frac{IQR}{\sqrt{n}}$, where IQR is the interquartile range.

Table 2.8: ESE motif density comparisons for SE and RI events in PSI clusters 1 and 2. One-tailed t-test p-values are shown. Alternative hypothesis is that the indicated binding motif density is greater in AS events that increase in PSI upon T3 treatment. NS indicates that the null hypothesis was not rejected at a significance level of 0.05.

AS event type	SRSF1	SRSF2	SRSF5	SRSF6
SE	$1.01e^{-06}$	0.00161	0.0287	NS
RI	0.00379	0.0379	NS	0.0013

The observation that ESE density correlates with splicing response demonstrates that the number of SR protein binding motifs is an important indicator of whether an alternatively included region of RNA will be present in the final transcript. RI and SE events appear to rely on different sets of SR proteins for their inclusion. Both SRSF1 and SRSF2 were predictive of splicing response in SE and RI events; However, SRSF5 was only predictive of response in SE events, and likewise SRSF6 for RI events.

2.3 CLK Inhibition Promotes Conjoined Gene Transcription in a Dose Dependent Manner

Inspection of splicing patterns using the Integrative Genomics Viewer [66] revealed cases of splicing between consecutive genes located on the same genomic strand in treated RNA-Seq libraries (Figure 2.16). Conjoined genes (CGs) have been previously reported in the literature, and are believed to arise from transcriptional read-through from the upstream to the downstream partner gene [67]. This hypothesis is supported by a common pattern: the second-to-last exon of the upstream gene being spliced to the second exon of the downstream gene. Skipping of the last and first exons of CG partner genes may be due to a lack of splicing signals at what would normally be a polyadenylation site or transcription start site, respectively. Additionally, the existence of intergenic exons in some CG transcripts strongly points to transcriptional read-through as the underlying mechanism for CG formation. Both of these patterns are present in the CGs detected in T3-treated samples.

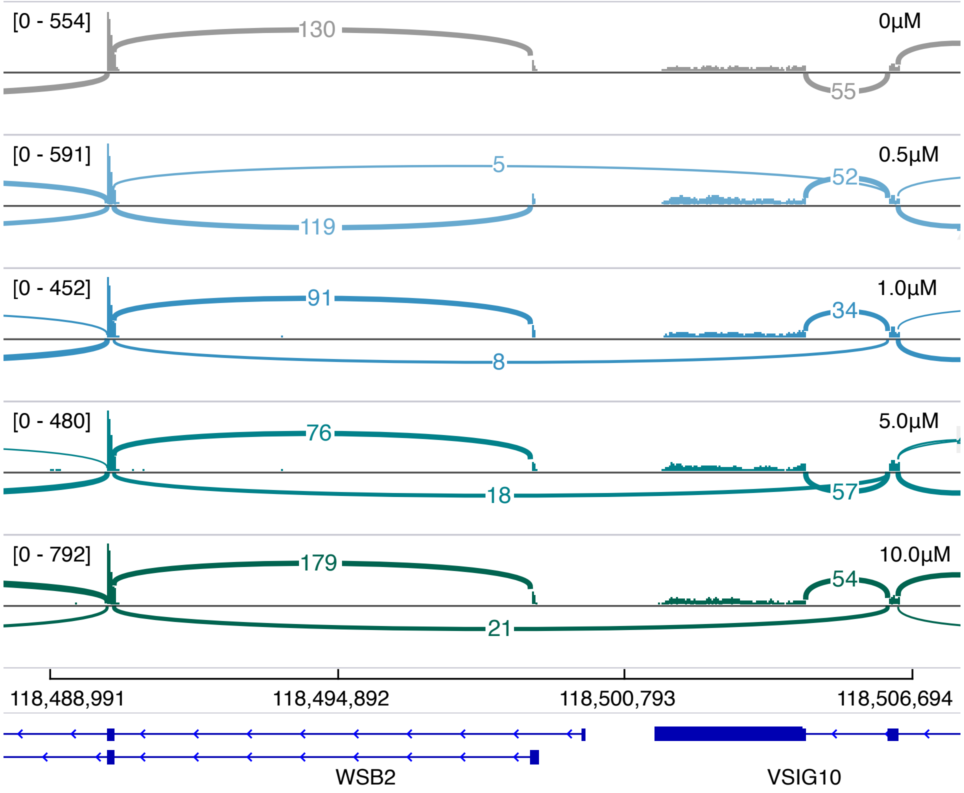

Figure 2.16: IGV-generated plot of splicing in the VSIG10-WSB2 conjoined gene. Plots for T3 treatment concentrations of 0.0, 0.5, 1.0, 5.0, and 10.0µM are shown from top to bottom. The control sample plot is coloured grey, and the treated sample plots are coloured according to T3 concentration. RefSeq gene annotations are shown in blue at the bottom of the plot along with chromosome 12 coordinates. For each sample, the y-axis represents read coverage, and the value range is indicated between brackets. Arcs connecting exons represent reads spliced across introns, with the number of spliced reads annotated over the line. Only arcs representing at least 3 reads are shown.

Although CLKs are known to play an important role in RNA splicing, the manner in which they might regulate 3′-end cleavage is unknown. Characterisation of CGs produced as a result of CLK inhibition is a first step towards understanding the genesis of these transcripts. Systematic analysis may provide insight into the regulation of 3′-end processing and reveal a novel role of CLK phosphorylation.

2.3.1 T3 treatment increases conjoined gene loci detection in a dose-dependent manner

A genome-wide search for further occurrences of conjoined transcripts was performed using the deFuse gene fusion detection method [55]. The deFuse classifier was modified by removing two features to increase CG detection sensitivity:

- est_breakseqs_percident
- breakseqs_estislands_percident

Conjoined genes are required to have both participating genes located on the same strand of the same chromosome. Detected CG events were filtered to have the following attributes:

- deletion = 'Y'
- expression $\geq$ 50 reads for both genes
- splice_score = 4 OR exonboundaries = 'Y'
- probability $\geq$ 0.9

These filters were chosen to produce a set of conjoined gene event calls that are likely due to splicing as opposed to genomic aberrations, and occur with a high probability. While it is likely that some real conjoined gene events have been missed due to stringent filter thresholds, this is acceptable as the focus of downstream analysis is on the characterisation of a set of true events rather than identifying all possible events.

Analysis of the RNA-Seq libraries in each T3-treated dataset revealed a common pattern of T3 dose-dependent detection of CG events (Figure 2.17). The HCT116 unstranded RNA-Seq dataset demonstrates a pattern similar to some AS events (*e.g.* SE) where the number of affected events increases dramatically at the 0.50µM T3 concentration. This pattern was not observed in the stranded RNA-Seq datasets due to the lack of measurements at T3 concentrations lower than 0.5µM. Nevertheless, the stranded RNA-Seq datasets do not contradict the results from the unstranded HCT116 dataset as they still reveal an increase in conjoined gene events at 0.5µM, with a milder dose effect. The similarity with AS events in dose-dependent response, especially the increase in event detection at 0.5µM, suggests that the production of CGs due to CLK inhibition is a primary effect of the treatment itself, rather than a secondary effect induced by disruption of the transcriptomic landscape.

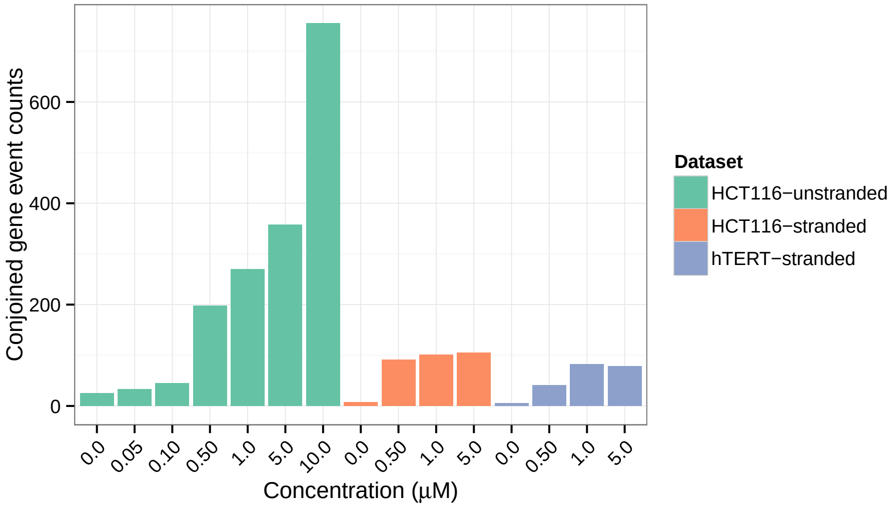

Figure 2.17: Conjoined gene counts per RNA-Seq library as detected by a modified deFuse classifier.

A substantial difference exists in the number of detected CGs between the unstranded and stranded RNA-Seq datasets (HCT116 unstranded RNA-Seq: 586, HCT116 stranded RNA-Seq: 215, hTERT: 154 unique events for 0.0µM, 0.5µM, 1.0µM, and 5.0µM; 2.7 fold increase in unstranded vs. stranded HCT116 RNA-Seq). This pattern was also observed in the number of differentially spliced AS events, and may be due to differences in the amount of splicing information in the RNA-Seq libraries from the stranded and unstranded RNA-Seq datasets (see Section 2.5).

Conjoined genes were also detected in RNA-Seq libraries generated from HCT116 cells transfected with CLK siRNA. 33 CGs (upstream, downstream gene pairs) were detected in the siRNA dataset after subtraction of CGs found in the control libraries. 25 of these CG were also found in the CG lists generated from the T3-treated sample libraries. Therefore, increased CG transcription can be explained by loss of CLK activity (as opposed to a T3 off-target), for at least some loci.

2.3.2 T3 treatment increases conjoined gene PSI in a dose-dependent manner

Increased detection of CG events upon T3 treatment implies a growth in CG transcription rate. If a constant fraction of transcripts from the upstream partner gene read through to the downstream partner, then increased CG transcription may indicate increased expression of the upstream partner gene. Alternatively, CLK inhibition may increase the proportion of transcripts escaping $3'$-end cleavage.

To investigate the affect of CLK inhibition on CG production rate, CG isoform annotations were generated and input into MISO. In cases where the second to last exon of the upstream CG partner gene is spliced to the second exon of the downstream partner, the isoform annotations can be constructed by using the last two exons of the upstream parent as the exclusion (wildtype) isoform, and the second to last exon of the upstream parent and the second exon of the downstream parent as the inclusion (CG) isoform. Any intergenic exons detected in CGs transcripts are included in the inclusion isoform annotations. When splicing occurs from the last exon of the upstream gene, annotations are generated where the terminal exon of the upstream parent is the exclusion isoform, and the same exon plus the appropriate exon of the downstream parent is the inclusion isoform. This class of annotations is similar to tandem UTR AS events in the MISO annotations. The generated CG isoform annotations were used by MISO to calculate PSI values for each CG event.

CGs were called as "differentially spliced" if MISO reported a Bayes factor $\geq$ 20, and a PSI difference $\geq$ 0.1 between treated and untreated samples. 603, 194, and 185 CGs were differentially spliced in the HCT116 unstranded RNA-Seq, HCT116 stranded RNA-Seq, and hTERT datasets. PSI value differences across all T3 concentrations were collected for each differentially spliced CGs, and CGs with missing PSI estimations were removed. PSI value distributions were then compared across each treatment concentration (Figure 2.18, Figure 2.19, Figure 2.20). Both the HCT116 and hTERT datasets show a dose-dependent increase in CG PSI (HCT116 unstranded RNA-Seq medians: 0.01, 0.07, and 0.14, for 0.10μM, 0.50μM, and 1.0μM; hTERT medians: 0.1, 0.16, and 0.23, for 0.50μM, 1.0μM, and 5.0μM). In the HCT116 unstranded RNA-Seq dataset, a clear increase (7 fold greater median

PSI than 0.10μM) in PSI value changes can be seen at the 0.5μM concentration. CLK inhibition clearly increases the proportion of CG to wild-type transcripts in a dose-dependent manner. CG PSI changes were then compared to the expression of non-conjoined upstream transcripts, and found that upstream gene non-conjoined transcription decreased with increased CG PSI (Figure 2.21, Figure 2.22, Figure 2.23). This pattern demonstrates that CGs "steal" transcription from the upstream CG participant.

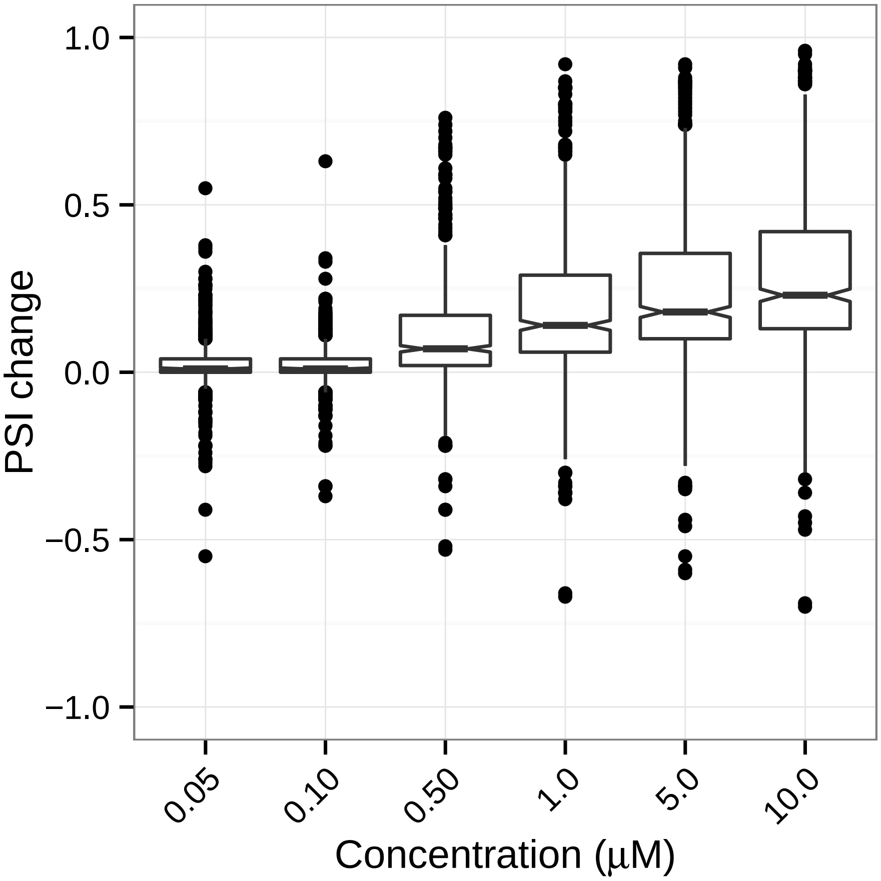

Figure 2.18: CG PSI change boxplots per T3 treatment for the HCT116 unstranded RNA-Seq dataset. $N = 603$. Notches extend $\pm 1.58\frac{IQR}{\sqrt{n}}$, where IQR is the inter-quartile range.

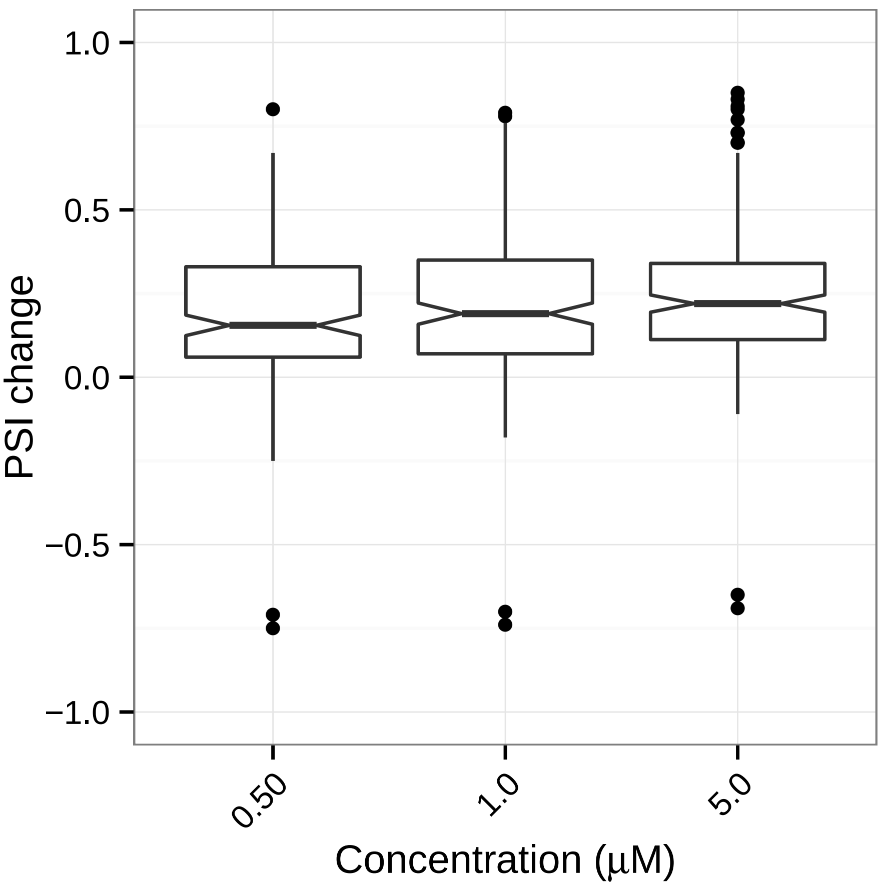

Figure 2.19: CG PSI change boxplots per T3 treatment for the HCT116 stranded RNA-Seq dataset. $N = 194$. Notches extend $\pm 1.58 \frac{IQR}{\sqrt{n}}$, where IQR is the inter-quartile range.

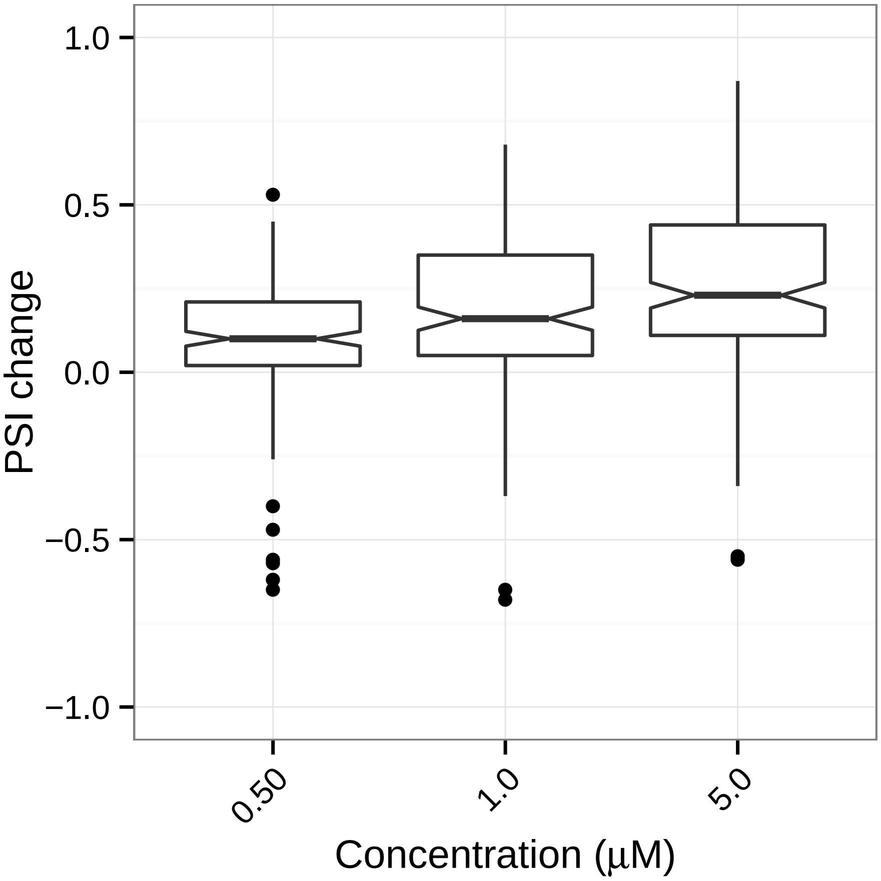

Figure 2.20: CG PSI change boxplots per T3 treatment for the hTERT stranded RNA-Seq dataset. $N = 185$. Notches extend $\pm 1.58 \frac{IQR}{\sqrt{n}}$, where IQR is the inter-quartile range.

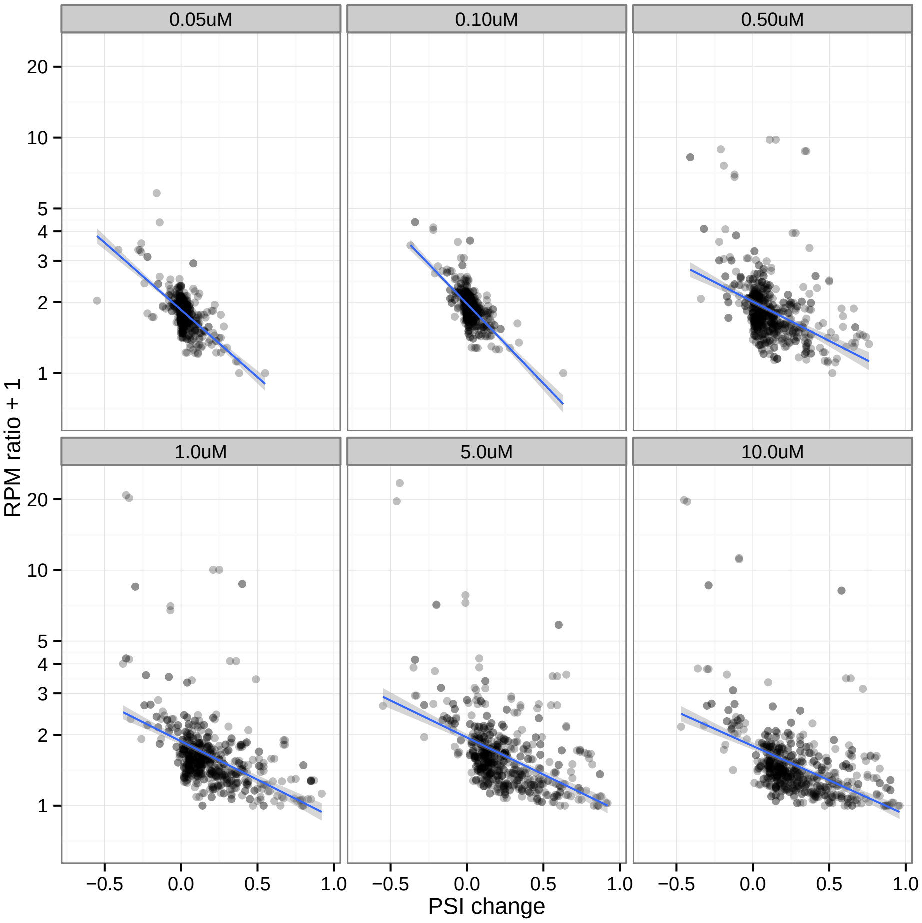

Figure 2.21: Non-conjoined upstream transcript expression ratio vs CG PSI change in the HCT116 unstranded RNA-Seq dataset. Upstream non-conjoined transcript expression is reads per million (RPM) mapped reads supporting the non-conjoined isoform from the CG MISO analysis. RPM ratio is the RPM of the upstream gene in the treated sample divided by the RPM in the control sample. PSI change is the difference in PSI from the control sample to the treated sample for the CG event. Negative regression line slope indicates decrease in non-conjoined transcription with CG PSI increase.

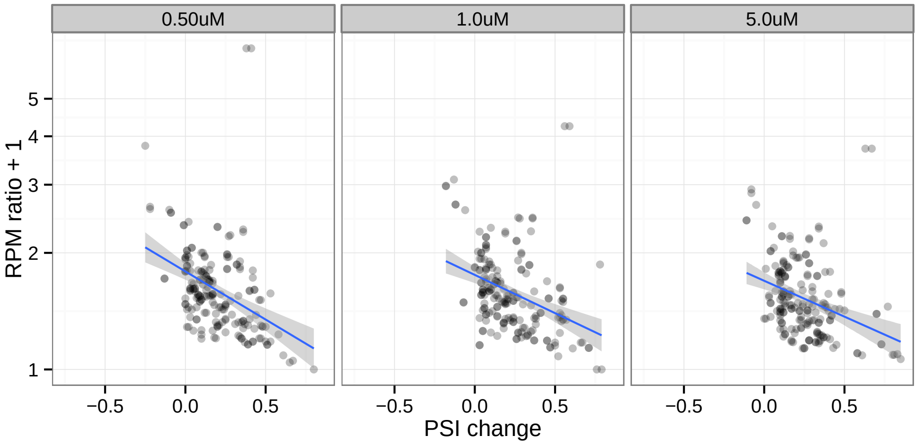

Figure 2.22: Non-conjoined upstream transcript expression ratio vs CG PSI change in the HCT116 stranded RNA-Seq dataset. Upstream non-conjoined transcript expression is reads per million (RPM) mapped reads supporting the non-conjoined isoform from the CG MISO analysis. RPM ratio is the RPM of the upstream gene in the treated sample divided by the RPM in the control sample. PSI change is the difference in PSI from the control sample to the treated sample for the CG isoform. Negative regression line slope indicates decrease in non-conjoined transcription with CG PSI increase.

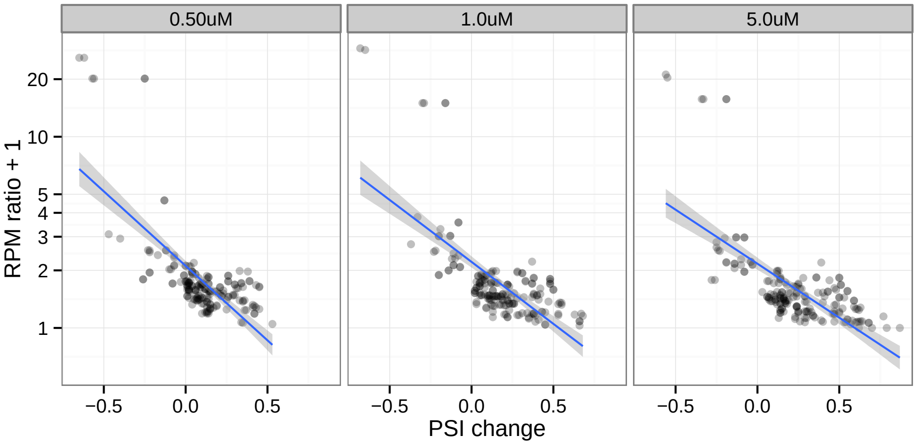

Figure 2.23: Non-conjoined upstream transcript expression ratio vs CG PSI change in the hTERT dataset. Upstream non-conjoined transcript expression is reads per million (RPM) mapped reads supporting the non-conjoined isoform from the CG MISO analysis. RPM ratio is the RPM of the upstream gene in the treated sample divided by the RPM in the control sample. PSI change is the difference in PSI from the control sample to the treated sample for the CG isoform. Negative regression line slope indicates decrease in non-conjoined transcription with CG PSI increase.

2.3.3 Similar conjoined genes are sensitive to CLK inhibition in HCT116 and hTERT cells

RNA 3′-end processing is differentially regulated according to cell type and tumour/normal status, similar to RNA splicing [36]. Despite differences in RNA processing regulation, CLK inhibition increases CG transcription in both malignant HCT116 and normal hTERT cells. Nevertheless, there may be differences in the set of CG loci between cell types. The degree of overlap between CGs will reflect the level of reliance on biological context in the vulnerability of genes to skip 3′-end cleavage.

Overlapping CGs were identified by generating a unique list of upstream-downstream CG partner pairs for each dataset. These lists ignore variation in donor and acceptor splice sites from the same CG partners, as these differences can be considered to arise from different isoforms (or "events") of the same CG. These lists were used to determine the set of conjoined genes common between cell types and exclusive to each cell type (Figure 2.24). 15 of 117 (12.8%) hTERT conjoined gene calls were not present in the HCT116 conjoined gene lists. Only 9 of 161 (5.6%) calls from the HCT116 stranded RNA-Seq dataset were not present in the other conjoined gene lists. Overall, the majority of both stranded RNA-Seq dataset CGs overlap with those in another dataset. However, 403 of 589 (68.4%) CGs called in the unstranded HCT116 RNA-Seq dataset were exclusive to that dataset.

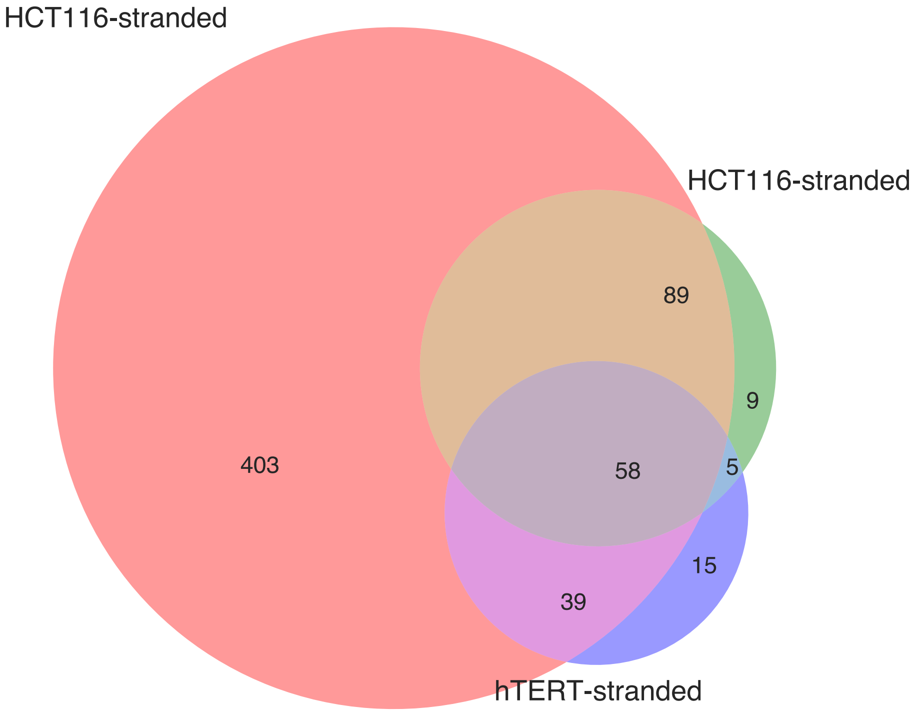

Figure 2.24: Venn diagram of conjoined genes detected in the two HCT116 and one hTERT RNA-Seq datasets.

The stranded RNA-Seq datasets both revealed many fewer CGs than the unstranded HCT116 RNA-Seq dataset. This is likely due, in part, to a greater number of treament concentrations in the unstranded RNA-Seq dataset, including the highest tested concentration (10.0μM). The majority of CGs in both of the stranded RNA-Seq datasets were detected in the unstranded RNA-Seq dataset. Also, large proportions of the HCT116 (55.3%) and hTERT (33.3%) stranded RNA-Seq datasets were detected in the unstranded RNA-Seq dataset, but not the other stranded RNA-Seq dataset. The unstranded RNA-Seq protocol may also be more sensitive to the detection of spliced sequences (see Section 2.5).

15 conjoined genes were only present in the high-confidence hTERT calls. The stringent filtering process for CG events may have removed the CG predictions found in the hTERT cells from the HCT116 predictions. Or, those particular CGs

may not have been sampled in the HCT116 RNA-Seq libraries, which could occur if expression is low.

CGs were also detected in HCT116 cells using cDNA sequences generated with the Pacific Biosciences' (PacBio) SMRT sequencing technology [9] (Table 2.1). The PacBio technology provides the ability to sequence up to several thousand nucleotides, allowing the capture of entire transcript sequences in many cases. As deFuse was designed to use paired-end RNA-Seq reads, an alternative method for CG detection was necessary for the PacBio data. Conjoined transcripts were detected by selecting reads that mapped across two different genes located on the same chromosome strand. For a PacBio read to be considered as "mapped" to a gene for the purposes of CG detection, at least three exon junctions within a read must match exon junctions belonging to a single gene in the Gencode level 1 and 2 transcript annotations. Cases where one gene is encapsulated within another gene (e.g. a miRNA located within the intron of another gene) are not considered conjoined genes. The result is an inclusive list of candidate CGs that can be compared to the CGs detected the RNA-Seq datasets.

The PacBio CGs were compared to those detected in the RNA-Seq datasets and overlapping CGs were counted in an identical manner to the RNA-Seq dataset comparison (Figure 2.25). 173 of 647 (26.7%) CGs detected in the PacBio dataset overlapped with those found in the RNA-Seq datasets. The PacBio-only CGs may be due, in part, to increased numbers of false positives: the detection method for the PacBio dataset was designed to favour sensitivity over specificity. However, the lower number of reads in the PacBio dataset (mean: 1,919,728) compared with the RNA-Seq datasets (*e.g.* HCT116 unstranded RNA-Seq mean: 167,167,942; approx. 87 times more than PacBio) means that the PacBio dataset may have sampled fewer conjoined gene transcripts. This may partially explain the lower overlap of the PacBio CGs with the RNA-Seq CGs.

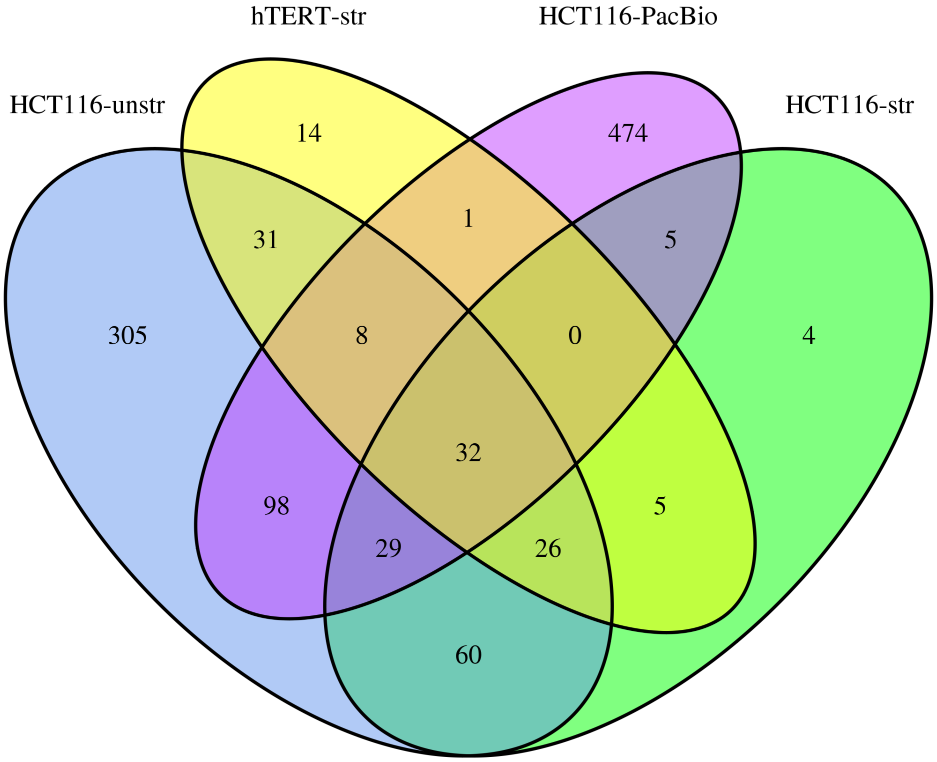

Figure 2.25: Venn diagram of conjoined genes detected in the RNA-Seq and PacBio datasets.

Only 1 of the 15 hTERT-specific CGs from the RNA-Seq dataset comparison was detected in the PacBio data. These hTERT-specific CGs may indicate a differential 3′ end processing response to CLK inhibition. These CGs may also be explained by cell-type specific gene expression profiles. Specifically, the hTERT-specific CGs may not be detected in the HCT116 samples merely due to low expression of the parent genes in HCT116 cells. To investigate this, FPKM values for genes involved in hTERT-specific CGs were calculated using Cufflinks for each of the HCT116 and hTERT datasets. The FPKM distributions of hTERT-specific CG partner genes reveal a pattern of higher expression in hTERT samples (Figure A.4, Figure A.5). Therefore, the presence of hTERT-specific CGs may be at least partially explained by reduced expression of participating genes in HCT116 cells.

2.3.4 Conjoined gene events are validated in both HCT116 and hTERT using targeted sequencing

While the PacBio dataset adds support for the presence of CGs detected in the RNA-Seq dataset, the low throughput and resulting lower sensitivity of the PacBio platform compared to RNA-Seq means that another validation method is necessary to properly estimate the proportion of true CG events. A set of 52 conjoined gene events (*i.e.* CG isoforms) was selected for targeted sequencing. The list of CGs include events found in both HCT116 and hTERT cells, and events found only in the CG lists of one cell type. The final list of sequencing amplicons also include regions of constitutive exons from three housekeeping genes. Housekeeping gene exon expression was used to normalize expression of each CG.

Targeted sequencing of the CG and housekeeping gene amplicons was performed on three datasets. Samples from the two HCT116 concentration curve experiments sequenced with unstranded and stranded RNA-Seq were used as two HCT116 replicate datasets. The hTERT concentration curve experiment samples were also used for CG targeted sequencing.

The validation sequencing libraries were analyzed for conjoined genes with deFuse. Detected CGs were compared to the set of CGs selected for validation. 37 of 52 (71.2%) CG events were validated with this method. Interestingly, 5 events not selected for validation were detected in the validation dataset. Upon inspection, 4 appear to be alternative isoforms of other CGs selected for validation; the other is similar to another validation input event except that it involves a more distant paralog of the upstream gene. This CG event is likely due to reads misaligned to the paralog gene. Considering CG parent genes only, and ignoring specific splice sites, 40 (76.9%) of the selected CGs were detected in the validation dataset.

Since the CGs chosen for validation include those found in only HCT116 or hTERT cells according to the deFuse analysis, the CGs found in the validation dataset may support the existence of cell-type specific CG events. However, of the 42 detected events, only 1 event was found in only one cell type — HCT116. One event was detected in a single HCT116 dataset and the hTERT dataset. The vast majority (40) of the detected events were present in all three (2 HCT116, 1 hTERT) datasets.

Many of the CGs detected in the validation dataset (22 of 42, or 52.4%) were also detected in untreated samples. To verify the effect of CLK inhibition on CG formation, CG event expression was compared across T3 treatment concentrations (Figure A.6, Figure A.7, Figure A.8). CG expression distributions show a dose-dependent increase in both HCT116 and hTERT cells.

2.3.5 Upstream partners of conjoined genes are involved in RNA metabolism and cell-cycle regulation

CG regulation may be focussed on either the upstream or downstream gene partners. For example, The downstream partners may use the promoter of the upstream gene to increase expression; Alternatively, CGs may form to add the downstream gene's functionality to the upstream gene. To investigate the possibility that upstream and downstream CG partners are involved in similar biological processes, the upstream and downstream partners were used to create two gene interaction networks using the ReactomeFI Cytoscape plugin [56]. The interaction network genes were checked for enriched GO biological process gene sets with false discovery rate controlled at 0.05. Enriched biological processes in the upstream and downstream CG partners were then used to generate an enrichment map [59] (Figure 2.26, Figure 2.27, Figure 2.28).

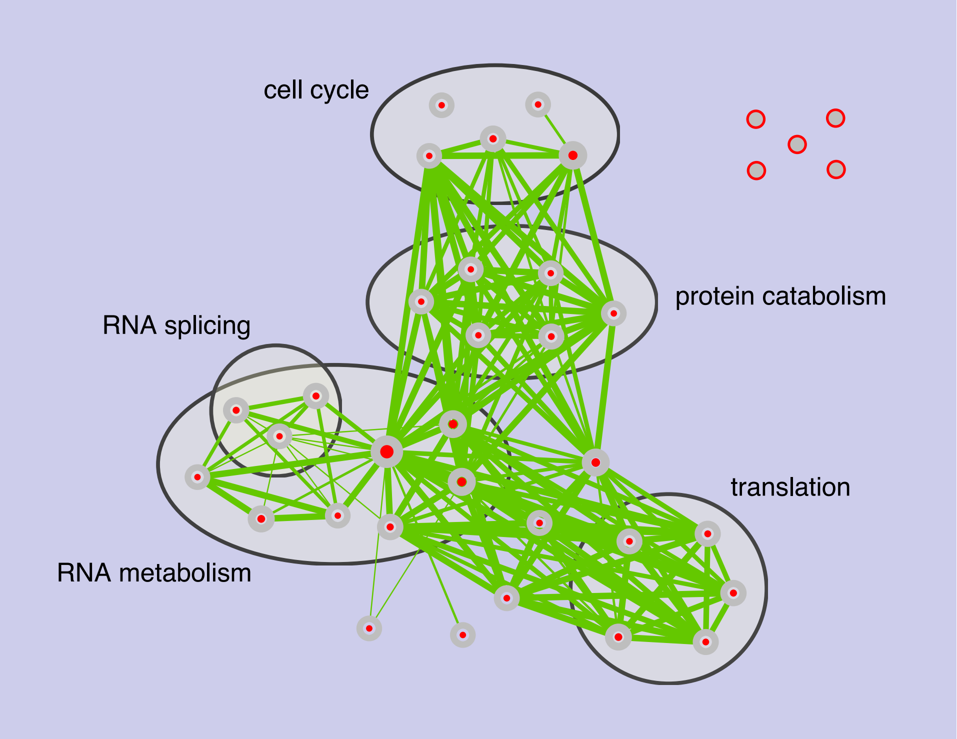

Figure 2.26: Enrichment map for genes involved in CGs in the HCT116 unstranded RNA-Seq dataset. Each node represents a GO biological process gene set. Biological processes enriched in CG upstream partners have red cores, while biological processes enriched in downstream partners have red outer rings. Edge thickness indicates the level of CG partner overlap between gene sets.

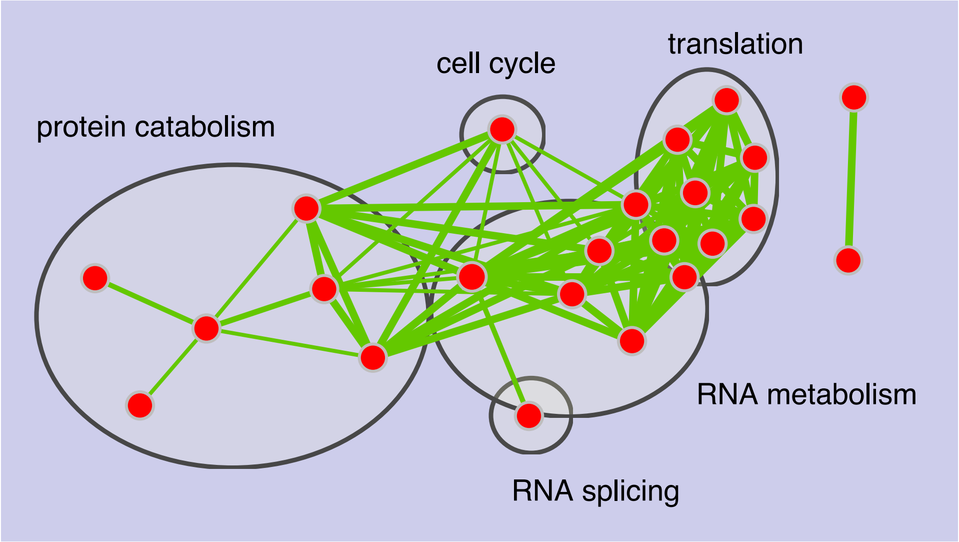

Figure 2.27: Enrichment map for genes involved in CGs in the HCT116 stranded RNA-Seq dataset. Each node represents a GO biological process gene set. Biological processes enriched in CG upstream partners have red cores, while biological processes enriched in downstream partners have red outer rings. Edge thickness indicates the level of CG partner overlap between gene sets.

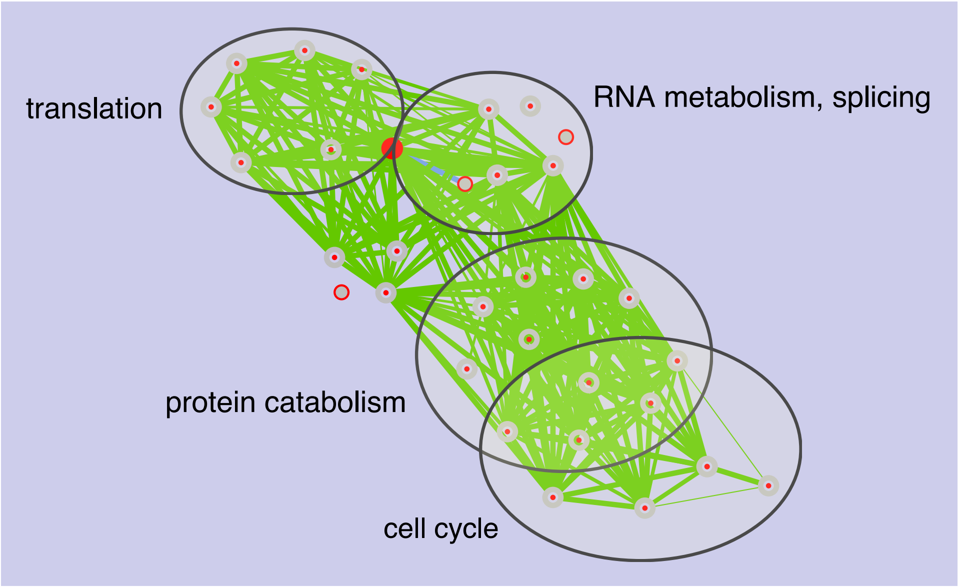

Figure 2.28: Enrichment map for genes involved in CGs in the hTERT dataset. Each node represents a GO biological process gene set. Biological processes enriched in CG upstream partners have red cores, while biological processes enriched in downstream partners have red outer rings. Edge thickness indicates the level of CG partner overlap between gene sets.

Analysis of the upstream gene partners produced a greater number of significantly enriched biological processes compared to the downstream partners. This suggests that the upstream gene partners are more related to each other, and that CG regulation is more focussed on the role of the upstream partners within the cell. Upstream CG partners are involved in RNA splicing, the cell cycle protein catabolism, and translation. Genes associated with $3'$-end processing, including A2AF1, CSTF1, and NUDT21 (a component of CFIm), were found to participate in CG transcription. Formation of CGs involving $3'$-end processing factors may disrupt normal $3'$-end cleavage, in turn promoting CG transcription at other loci.

Similar biological processes were affected by CG transcription (Figure 2.26, Figure 2.27, Figure 2.28) and differential splicing (Figure 2.3, Figure 2.4, Figure 2.5, Table 2.3), which may suggest that CLKs can regulate this common set of biological functions through different RNA processing mechanisms. Formation of CGs might comprise one aspect of cellular response to CLK inhibition. For example, CG transcription may be a mechanism for upstream gene expression control if the CG transcript is targeted for degradation by the nonsense-mediated decay pathway [67].

2.3.6 Upstream conjoined gene partners may rely on auxiliary $3'$-end processing factors

Transcriptional readthrough of upstream CG partners into downstream genes may be regulated by components of the $3'$-end processing machinery. While a gene may have multiple alternative cleavage and polyadenylation sites, CG formation requires the skipping of all possible sites in the upstream gene. Yet, final poly(A) sites generally contain a strong, canonical poly(A) signal [36]. Terminal poly(A)/$3'$ cleavage sites of upstream CG partners may contain common *cis*-regulatory signal patterns that are sensitive to RS domain phosphorylation status. These genes would then be susceptible to transcriptional read-through upon CLK inhibition.

Regulatory signals associated with CG formation were investigated by identifying the annotated locations of terminal poly(A) sites in the genome. The proportion of upstream CG partners with canonical poly(A) signals at their terminal poly(A) site was similar to the proportion for all genes. Therefore, the absence of a canonical poly(A) signal alone does not appear to be associated with CG generation. The regions around the terminal poly(A) sites were examined for the presence of a canonical A(A/U)UAAA poly(A) signal, an upstream UGUA signal, and a U/GU-rich downstream element (DSE). For the purposes of this analysis, a DSE is defined as a sequence of at least six nucleotides, composed of uracils and interspersed with up to three non-sequential guanines.

Polyadenylation sites without canonical poly(A) signals are known to rely on auxiliary 3′-end processing factors for poly(A) site selection [36]. So, genes were partitioned into two groups based on whether or not their terminal poly(A) site contained a nearby canonical poly(A) signal, as detected through this analysis. Upstream CG partners in the group lacking canonical poly(A) signals had a higher proportion of detected UGUA signals (chi-squared p-value < 0.01) and DSEs (chi-squared p-value < 0.05) compared to all genes without an annotated poly(A) site. This pattern was not found in a similar comparison with the group containing nearby canonical poly(A) signals.

Upstream CG gene partners lacking canonical poly(A) signals seem to rely on CFIm binding to UGUA sites and CstF binding to G/GU-rich DSEs more often than typical genes. Proper 3′ cleavage of these genes may be especially sensitive to regulation of CFIm and CstF. The heavier reliance on CFIm binding in particular is interesting, because SR proteins are known to interact with CFIm, potentially by assisting in the recruitment of CFIm to the RNA substrate [25]. Furthermore, phosphorylation of CFIm is necessary for the 3′ cleavage reaction to occur [26]. CLKs may regulate RS domain mediated SR protein-CFIm interactions, or may even phosphorylate the RS-like domain of CFIm itself. This may partially explain the sensitiviy of these CG loci to CLK inhibition. For genes with canonical poly(A) signals at terminal polyadenylation sites, CG formation propensity may be determined by regulation of core components of the 3′-end processing machinery.

2.4 CLK Inhibition Results in the Down Regulation of Splicing Factors and Cell Cycle Regulators

CLK inhibition causes widespread structural changes in the transcriptome. Any gene expression changes could be due to changes in transcriptome composition, or a direct response to the presence of the T3 compound. Cufflinks [51] was used to quantify transcript abundances in the three T3-treated RNA-Seq datasets, which produced FPKM values for each gene. Genes selected for further analysis were required to have FPKM values $>=1$ in at least 4 libraries for the unstranded HCT116 RNA-Seq dataset, and 3 libraries for the two stranded RNA-Seq datasets. This filtering was performed to remove unexpressed genes that have a low FPKM value due to the presence of misaligned reads. Each gene must also have an FPKM fold change $>=2$ for at least one treated library when compared with the untreated control library. The resulting list represents candidate differentially expressed genes.

Determining whether a gene is differentially expressed in a statistically meaningful manner without biological replicates is challenging. However, by measuring RNA at a variety of CLK inhibitor concentrations, genes with expression profiles following clear trends across the concentration gradient can identified as likely to be differentially expressed. Gene expression trends were discovered by clustering gene expression profiles using the WGCNA [64] clustering method. WGCNA was run with networkType="signed", minModuleSize=25, and power=28 for the unstranded HCT116 RNA-Seq dataset, power=27 for the stranded HCT116 RNA-Seq dataset, and power=30 for the HTERT dataset. This resulted in 6 clusters for the unstranded HCT116 RNA-Seq dataset, 5 clusters each for the stranded HCT116 RNA-Seq and hTERT datasets. For each cluster, a representative gene expression profile (an "eigengene") was calculated and genes whose expression profiles correlated with the eigengene expression profile less than 0.75 were removed.

All three datasets exhibit similar FPKM profile clusters (Figure 2.29, Figure 2.30, Figure 2.31). Both stranded RNA-Seq datasets include fewer treatment libraries and so the expression profiles will appear somewhat different. The number of down-regulated genes greatly outnumbered up-regulated genes. The splicing and transcriptional machineries are linked and splicing disruption may have caused a negative effect on gene expression. In all datasets the largest cluster is characterised by genes that are strongly down-regulated starting at the 0.5μM concentration. Some clusters behave in an opposing manner: their genes are strongly up-regulated at the same concentrations. This pattern of greater regulatory activity at the 0.5µM concentration was also observed in the differential splicing analysis (Section 2.2), and the CG analysis (Section 2.3). The cluster profiles for both HCT116 datasets and the hTERT dataset demonstrates that gene regulatory processes are affected similarly in both HCT116 and hTERT cells as a result of CLK inhibitor treatment.

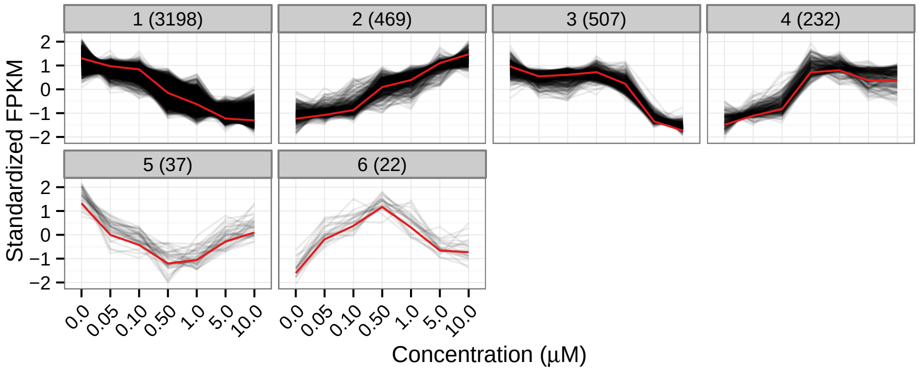

Figure 2.29: Clustered gene expression profiles from the HCT116 unstranded RNA-Seq dataset. Genes have been clustered using WGCNA based on FPKM profiles. Each black line is a gene expression profile; The red lines are cluster eigengenes.

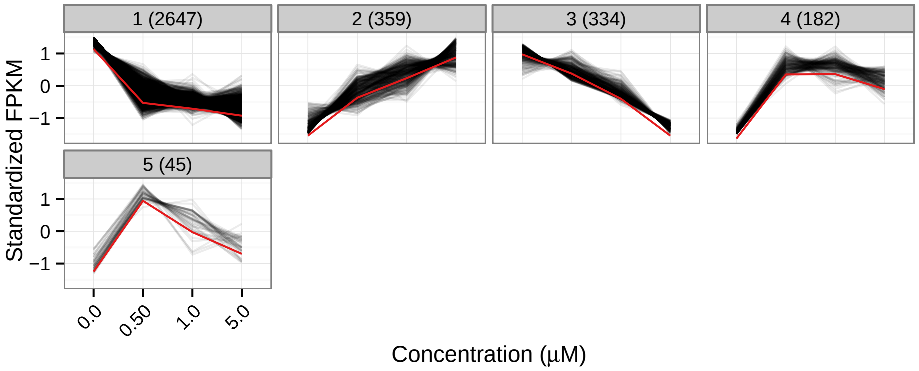

Figure 2.30: Clustered gene expression profiles from the HCT116 stranded RNA-Seq dataset. Genes have been clustered using WGCNA based on FPKM profiles. Genes have been clustered using WGCNA based on FPKM profiles. Each black line is a gene expression profile; The red lines are cluster eigengenes.

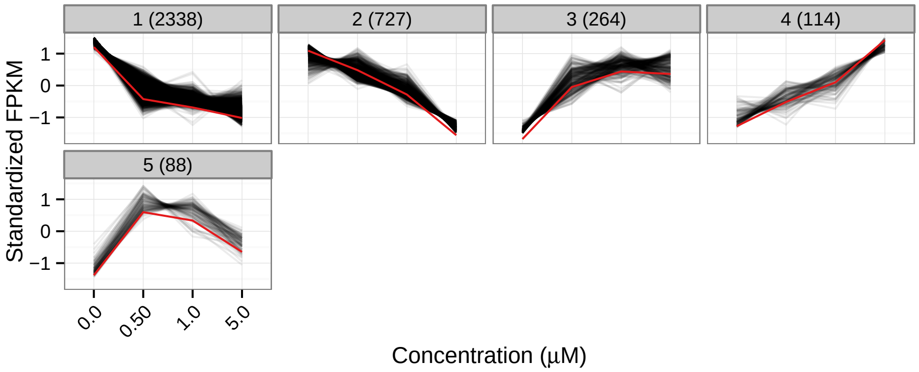

Figure 2.31: Clustered gene expression profiles from the hTERT stranded RNA-Seq dataset. Genes have been clustered using WGCNA based on FPKM profiles. Genes have been clustered using WGCNA based on FPKM profiles. Each black line is a gene expression profile; The red lines are cluster eigengenes.

Each cluster contains genes that appear to be subject to similar regulatory processes. Therefore, it is likely that each cluster contains groups of genes that participate in similar or related biological processes. Identifying biological processes enriched within each gene expression cluster will provide a glimpse into how biological processes are affected by differential expression due to CLK inhbition.

Functional enrichment analysis of clustered genes was performed using the ReactomeFI Cytoscape plugin [56]. For each set of clustered genes, a gene interaction network was constructed and genes remaining in the constructed network were used to perform functional enrichment analysis. Enriched GO biological process terms with false discovery rate controlled at 0.05 were reported for each cluster. For the HCT116 datasets, only analysis of clusters 1–3 resulted in a list of enriched biological processes. The hTERT dataset only produced enriched biological processes for clusters 1–4. Enriched biological processes were used to create enrichment maps [59] (Figure 2.32, Figure 2.33, Figure 2.34).

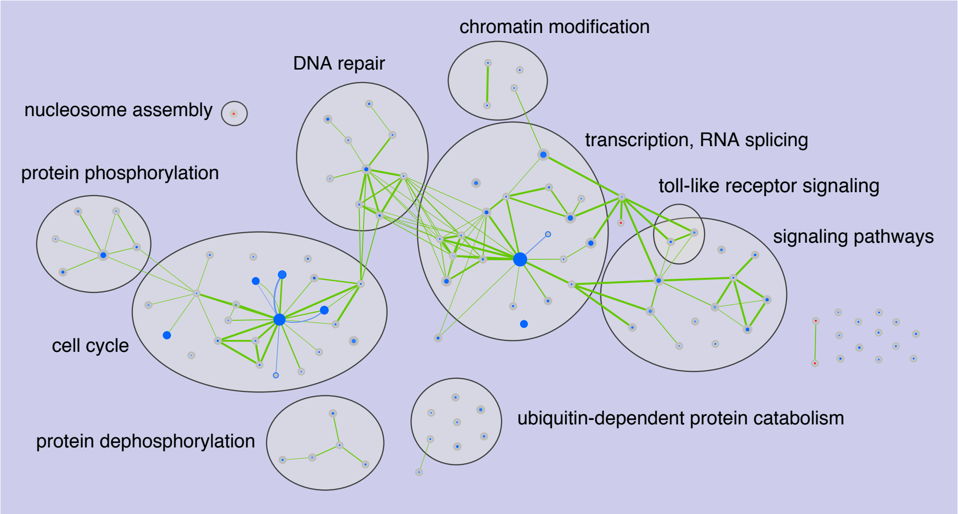
chromatin modification
DNA repair
nucleosome assembly
transcription, RNA splicing
protein phosphorylation
toll-like receptor signaling
signaling pathways
cell cycle
ubiquitin-dependent protein catabolism
protein dephosphorylation

Figure 2.32: Biological process enrichment map for differentially expressed genes in the HCT116 unstranded RNA-Seq dataset. Each node represents a GO biological process gene set. Red nodes represent biological processes enriched among up-regulated genes, likewise blue for down-regulated genes. Node cores are coloured blue when that gene set is enriched among genes in cluster 1, red for cluster 2. The outer ring is coloured blue when that gene set is enriched among genes in cluster 3. Edge thickness indicates the level of overlap between two gene sets, considering the set of up- or down-regulated genes.

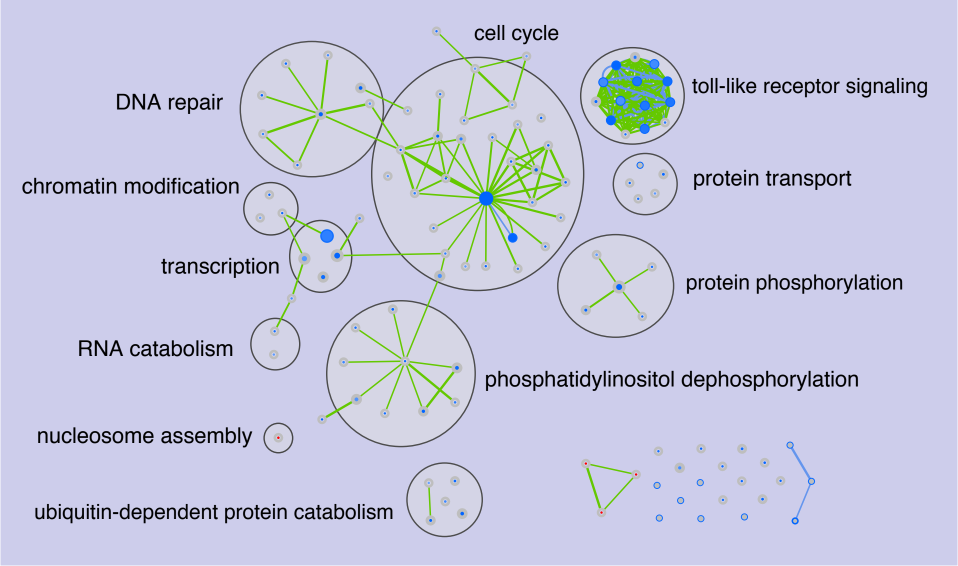

Figure 2.33: Biological process enrichment map for differentially expressed genes in the HCT116 stranded RNA-Seq dataset. Each node represents a GO biological process gene set. Red nodes represent biological processes enriched among up-regulated genes, likewise blue for down-regulated genes. Node cores are coloured blue when that gene set is enriched among genes in cluster 1, red for cluster 2. The outer ring is coloured blue when that gene set is enriched among genes in cluster 3. Edge thickness indicates the level of overlap between two gene sets, considering the set of up- or down-regulated genes.

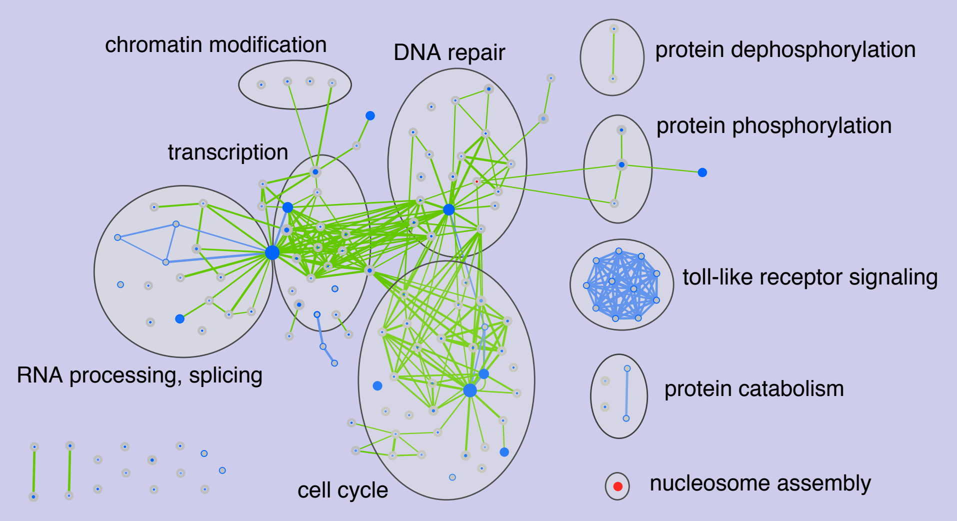

Figure 2.34: Biological process enrichment map for differentially expressed genes in the hTERT dataset. Each node represents a GO biological process gene set. Red nodes represent biological processes enriched among up-regulated genes, likewise blue for down-regulated genes. Node cores are coloured blue when that gene set is enriched among genes in cluster 1, red for cluster 3. The outer ring is coloured blue when that gene set is enriched among genes in cluster 2, red for cluster 4. Edge thickness indicates the level of overlap between two gene sets, considering the set of up- or down-regulated genes.

Genes characterised by strong down-regulation at the 0.5μM concentration (cluster 1) are enriched for RNA splicing and processing genes (Figure 2.32, Figure 2.33, Figure 2.34). Up-regulated gene clusters were not enriched for RNA splicing and processing genes. The down regulation of genes involved in RNA metabolism may represent an attempt by treated cells to prevent the production of aberrant RNA transcripts due to CLK inhibition.

Aside from RNA processing, genes involved in cell cycle regulation were down-regulated. Down-regulation of cell cycle regulators upon T3 treatment suggests that CLK inhibition may disrupt normal cell cycle activity. RNA splicing is inhibited during mitosis [61] and appears to involve the dephosphorylation of SRSF10 proteins [68]. In addition, down-regulation of SRSF3 induces G1 cell cycle arrest in HCT116 colon cancer cells [62]. Splicing repression via CLK inhibition may have a similar effect.

Genes in the second down-regulated cluster (3 for HCT116, 2 for hTERT) were fewer than those in cluster 1 and were enriched for many fewer biological process gene sets. Biological processes enriched in the secondary down-regulated cluster overlapped with those of cluster 1, and are related to RNA metabolism and cell cycle regulation. A subset of genes are perhaps more resilient to expression changes in the presence of CLK inhibition, and increasing T3 dose is progressively disrupting biological processes.

Toll-like receptor signaling genes were down-regulated upon CLK inhibition. However, this biological process seemed to be more sensitive in HCT116 cells than hTERT cells. In HCT116 cells, toll-like receptor signaling was down-regulated in cluster 1 (strong down-regulation at 0.5μM) as well as cluster 3 (more resilient to down-regulation). In hTERT cells, toll-like receptor signaling was down-regulated in cluster 2 (more resilient to down-regulation).

Up-regulated genes were much fewer than down-regulated genes and thus affected fewer biological processes. Histone assembly was among the few biological processes found to be enriched among only up-regulated gene expression clusters in all three datasets.

Biological processes affected by gene down-regulation are consistent with the biological processes affected by differential splicing and CG transcription. RNA metabolic processes (including splicing), cell cycle, and protein catabolism are affected by changes in all three processes. Both differential splicing and gene down-regulation affected DNA repair, histone modification, protein phosphorylation, and toll-like receptor signaling. SR proteins are reported to play a role in transcriptional elongation, and depletion of some SR proteins can have a negative impact on transcription [69]; Disruption of SR protein activity via CLK inhibition may attenuate the splicing and expression a common set of genes, potentially explaining the similarities in biological processes affected by differential splicing and expression down-regulation.

2.5 Comparison of Unstranded and Stranded RNA-Seq Libraries

The stranded RNA-Seq datasets produced many fewer significant AS and CG events (Figure 2.1, Figure 2.17). To identify sources of these differences, the RNA-Seq libraries were compared to each other and to the PacBio libraries using various metrics. First, PSI values for each AS event were compared at each common T3 concentration between the HCT116 unstranded and stranded RNA-Seq libraries (Figure 2.35). AS events were not compared at a certain T3 concentration if they did not pass a coverage threshold in both datasets of 1 read each for both the inclusion and exclusion isoforms and 10 reads total for the AS event. This read coverage filter is the same as applied for the MISO differential splicing analysis. The unstranded and stranded RNA-Seq dataset AS event PSI values had a Pearson correlation coefficient of 0.75. A pattern of anti-correlation amongst a subset of events was also observed.

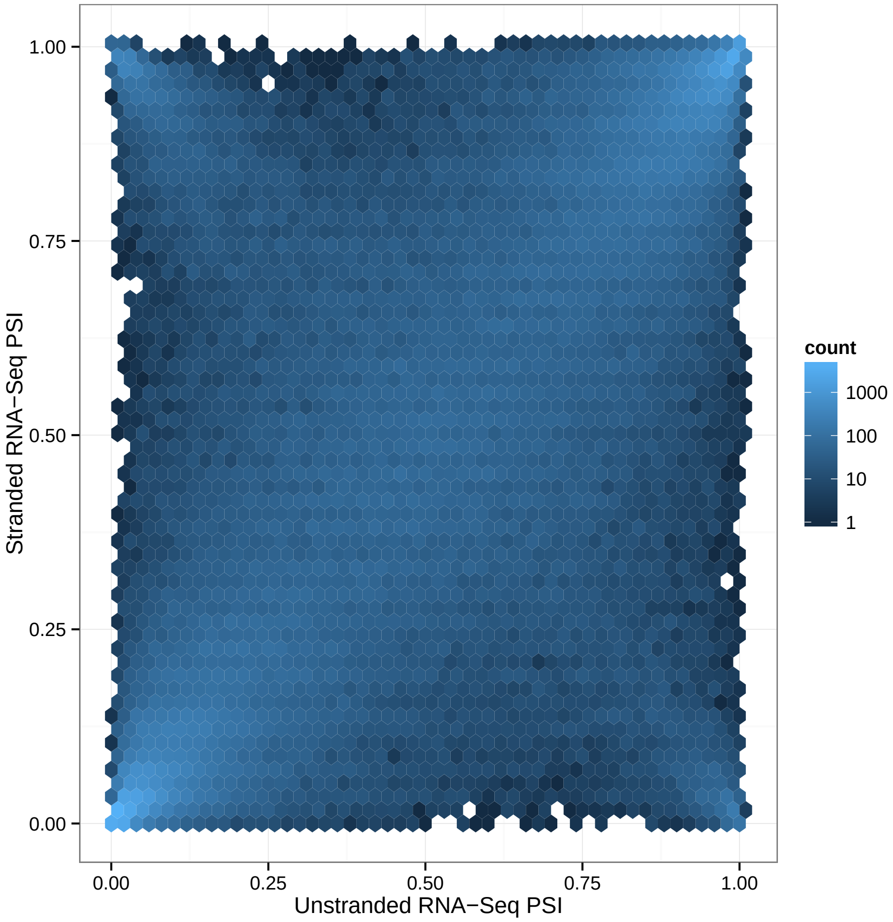

Figure 2.35: HCT116 unstranded vs. stranded RNA-Seq hexplot of AS event PSI values. PSI values were compared for each event at each concentration. Each hex represents a number of AS events. The lighter the shade of blue, the greater the number of AS events map to that hex.

Similarly, the unstranded and stranded HCT116 RNA-Seq AS event PSI values were compared to PSI values computed from the PacBio sequencing libraries (Figure 2.36, Figure 2.37). PacBio reads violate some assumptions of the MISO model, so PSI values were calculated by counting reads supporting the inclusion and exclusion isoforms in the MISO event annotations. PacBio PSI values were more strongly correlated with the unstranded RNA-Seq dataset (Pearson correlation coefficient 0.76) compared with the stranded RNA-Seq dataset (Pearson correlation coefficient 0.66). Anti-correlation can also be observed amongst a subset of events in the PacBio vs. RNA-Seq comparisons, although perhaps to a lesser extent in the comparison with the unstranded RNA-Seq data. The higher correlation between the PacBio PSI and unstranded RNA-Seq PSI values suggests that the unstranded RNA-Seq PSI values may be more reliable than those computed from the stranded RNA-Seq dataset.

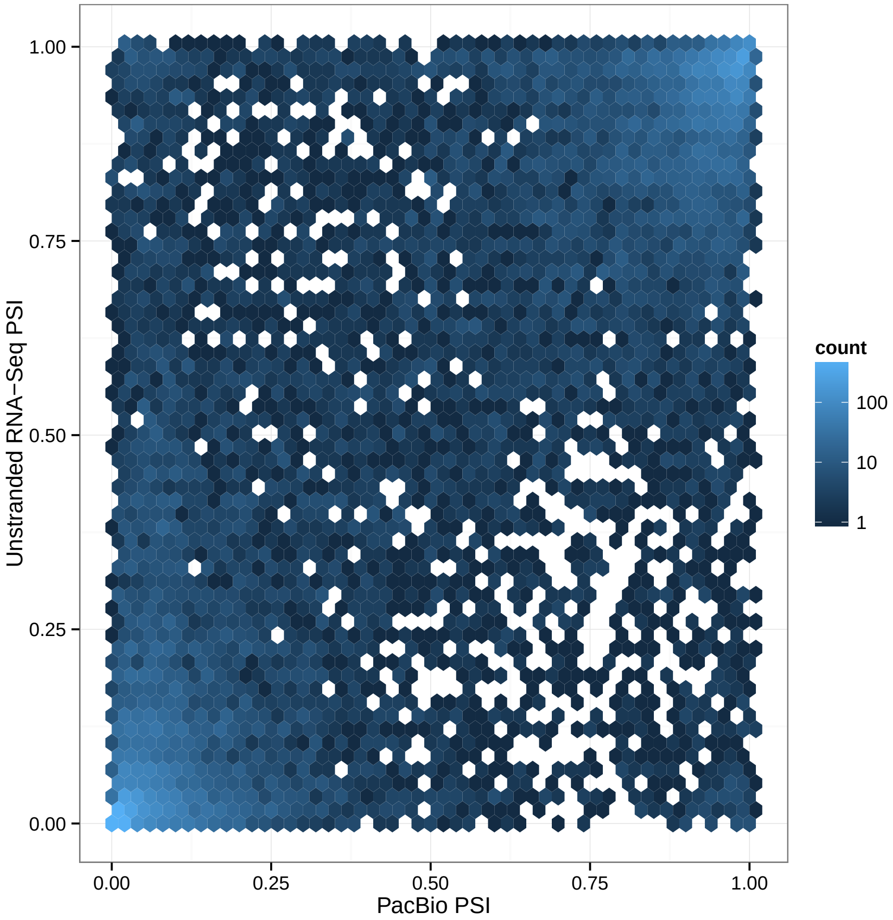

Figure 2.36: HCT116 PacBio vs. unstranded RNA-Seq hexplot of AS event PSI values. PSI values were compared for each event at each concentration. Each hex represents a number of AS events. The lighter the shade of blue, the greater the number of AS events map to that hex.

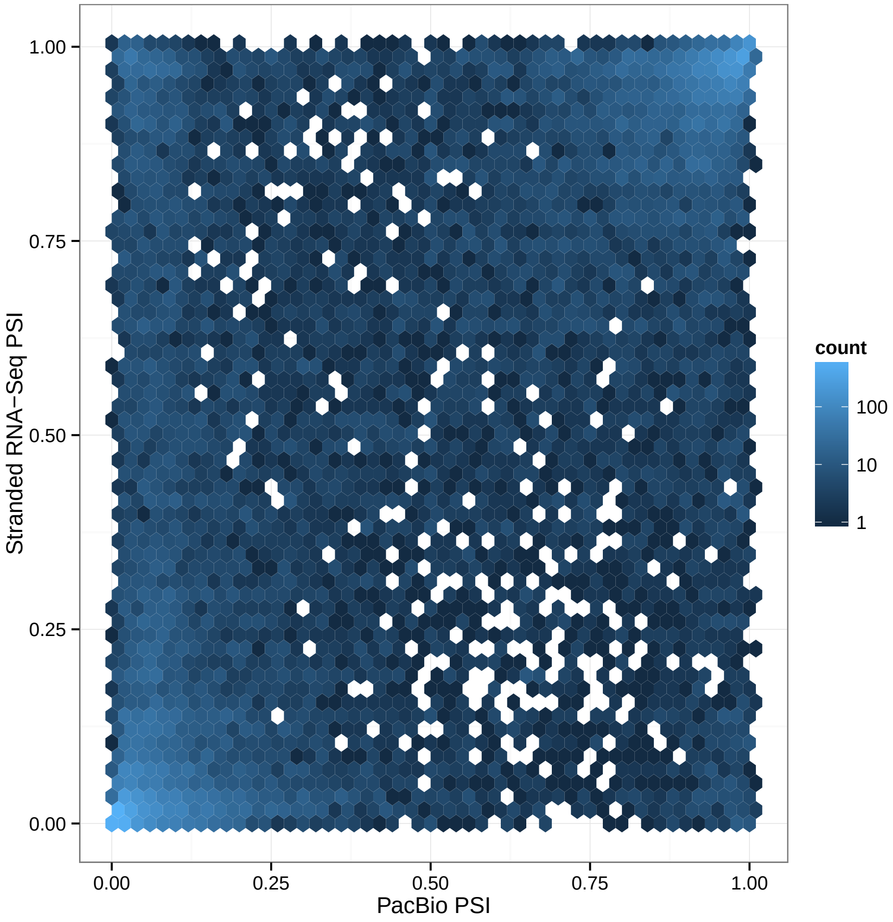

Figure 2.37: HCT116 PacBio vs. stranded RNA-Seq hexplot of AS event PSI values. PSI values were compared for each event at each concentration. Each hex represents a number of AS events. The lighter the shade of blue, the greater the number of AS events map to that hex.

Next, the number of mapped reads between the three RNA-Seq datasets were compared at each common T3 concentration (Figure 2.38). Generally, the unstranded RNA-Seq libraries have a greater number of mapped reads. However, this pattern is not always consistent; At the 1.0μM concentration the number of mapped reads is roughly equal between the three datasets. Therefore, while read coverage may play a role in event count differences between the unstranded and stranded RNA-Seq libraries, it cannot be the primary cause.

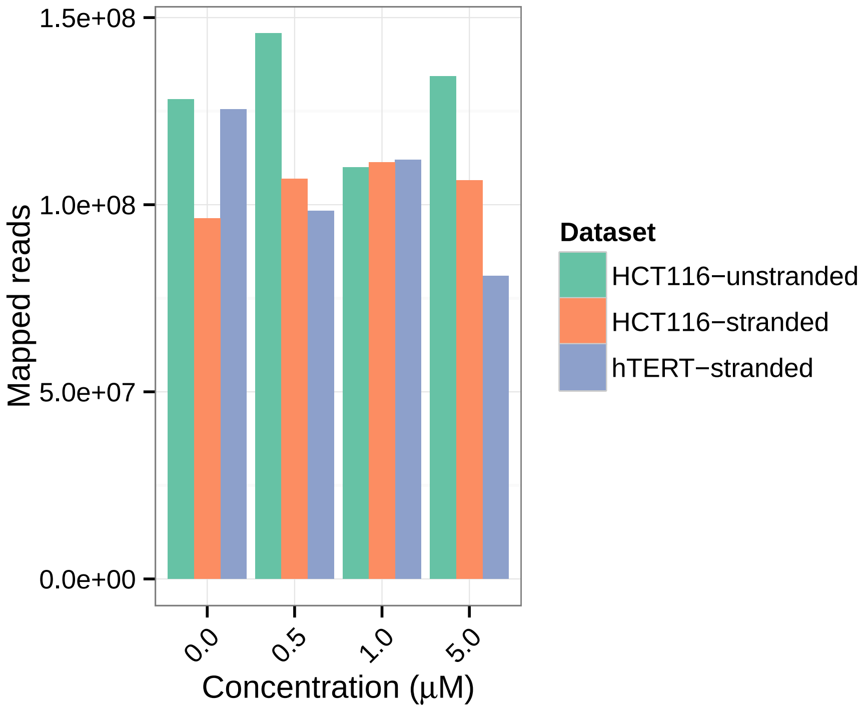

Figure 2.38: Mapped read counts for RNA-Seq libraries from the three T3-treated RNA-Seq datasets. Counts for T3 concentrations common amongst the three datasets are shown.

Finally, the proportion of mapped reads that were split during the mapping process were compared (Figure 2.39). The majority of these reads are split across introns and are an important source of evidence for RNA splicing in a sequencing library. A lower proportion of split reads may result in a reduced ability to detect and quantify alternative splicing. Lower split read proportions were detected in the stranded RNA-Seq libraries. In both HCT116 datasets the proportion of split reads decreases with increasing T3 dose. The hTERT dataset shows a similar dose-dependent effect, however the decrease in split read proportion is not as strong, especially at the higher concentrations. This weaker dose effect in the hTERT dataset can also be observed in the differentially spliced AS event and CG event counts (Figure 2.1, Figure 2.17). Differences in the proportion of mapped reads that align to splice junctions appears to be a main contributor to the reduction of detected splicing events in the stranded RNA-Seq datasets.

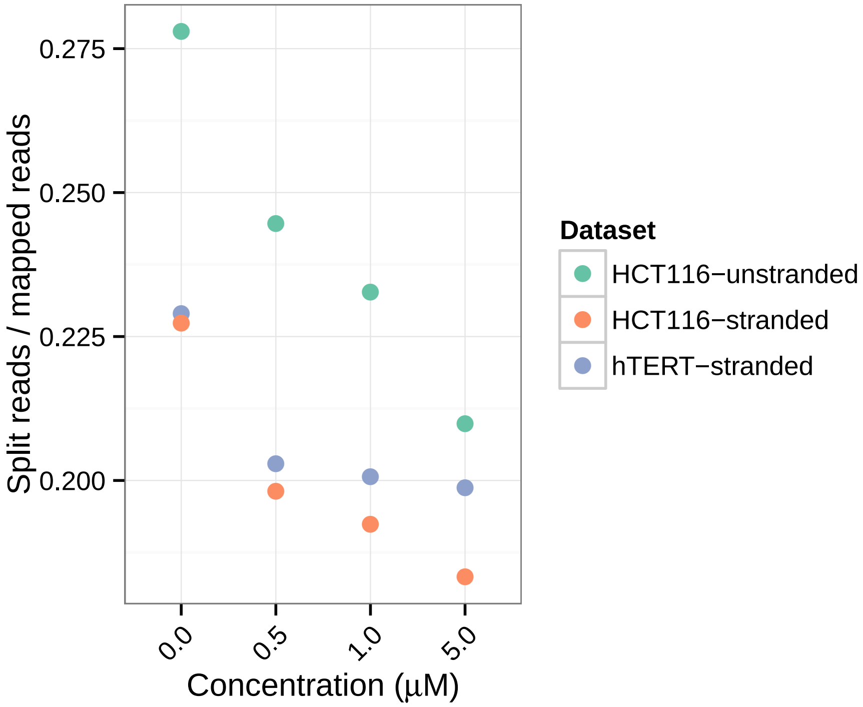

Figure 2.39: Proportion of mapped reads split during the alignment process. Proportions for T3 concentrations common amongst the three datasets are shown.

Dose dependent decreases in split read proportions may be explained by the increasing presence of aberrantly spliced transcripts. The GSNAP aligner may struggle to map splice junction reads from novel splice sites in these transcripts. The overall lower proportion of split read proportions in the stranded RNA-Seq libraries may suggest that the unstranded RNA-Seq libraries contain a greater proportion of reads erroneously mapped to non-contiguous regions of the genome. However, the higher correlation of unstranded RNA-Seq and PacBio AS event PSI values suggests that the opposite may be true: the stranded RNA-Seq datasets may include less RNA splicing information.

Chapter 3

Discussion

The results of the analyses presented in this thesis has demonstrated that the T3 CLK inhibitor is an effective disruptor of normal RNA processing. Applying the CLK inhibitor to cells in progressively greater concentrations allowed dose-dependent response patterns to be observed in alternative splicing regulation, $3'$-end processing (*i.e.* conjoined gene formation), and gene expression regulation. Performing concentration-curve experiments in both HCT116 colon cancer and normal hTERT cells revealed that the majority of observable effects on the transcriptome were not specific to cancer or normal biology.

AS events exhibited varying levels of sensitivity to CLK inhibition. For example, SE events displayed a sharp decrease in PSI starting at the 0.5μM concentration, compared to lower concentrations. RI events appear to be less dependent on CLK activity and began to show large increases in PSI at the 1.0μM concentration. These splicing responses clearly indicate that CLK inhibition disrupts splice site recognition.

The RS domain of SR proteins are generally thought to facilitate protein-protein interactions. However, a recent study has shown that phosphorylation is required for the RS domain of SRSF1 to dissociate from the RRM domain, allowing the RRM domain to recruit U1 snRNP [14]. Under either model, repressing RS domain phosphorylation prevents SR proteins already bound to the RNA substrate from promoting spliceosome formation. Therefore, RNA-bound and unphosphorylated SR proteins may directly inhibit splicing.

ESE density appears to be an important predictor of AS inclusion levels. Alternative sequences in SE and RI events that are up-regulated upon CLK inhibition tend to have a greater density of ESE motifs than down-regulated exons. Greater ESE density provides more opportunities for SR protein binding, and increases the chances of a sufficiently phosphorylated SR protein being available to recruit members of the spliceosome. SEs and RIs appear to be regulated by different SR proteins; SE events that decreased in PSI with treatment were depleted of SRSF1, SRSF2, and SRSF5 binding motifs. Similarly responding RI events were depleted of SRSF1, SRSF2, and SRSF6 binding motifs.

RNA 3′-end cleavage was also shown to be negatively impacted by CLK inhibition. Conjoined gene formation occurred in a T3 dose-dependent manner and, similar to SE events, greater effects were observed starting at 0.5μM. Targeted sequencing of a subset of detected CGs recapitulated these results, and verified the existence of detected CGs in untreated cells. Dose-dependent increases in CG PSI and decreases in non-conjoined upstream gene transcription indicate that CG expression is "stolen" from the upstream gene.

Conjoined gene formation through transcriptional read-through appears to be a natural phenomenon and has received some attention in the literature [67, 70]. T3-induced CG production patterns suggest that CLK phosphorylation is important for the 3′-end cleavage reaction of some genes. U2AF, a component of the spliceosome, has been shown to promote 3′-end cleavage by interacting with CFIm [20]. SR proteins facilitate the recruitment of U2AF to 3′ splice sites [11], and thus may indirectly promote recruitment of CFIm when properly phosphorylated. However, SR proteins have also been shown to interact directly with CFIm [25], and so may also directly promote its recruitment. Involvement of CFIm in CG transcription regulation is supported by the finding that, among genes lacking canonical poly(A) signals at their terminal polyadenylation site, upstream CG partners have a higher proportion of terminal polyadenylation sites with CFIm-binding UGUA signals. Interestingly, phosphorylation of CFIm is required for the 3′-end cleavage reaction to occur [26]. This presents the possibility that CLK phosphorylates the RS-like domain of CFIm itself and regulates 3′-end processing.

T3 treatment revealed 5–6 gene expression response patterns to CLK inhibition, with the bulk of genes being down-regulated upon treatment. For most differentially expressed genes, greater changes in expression were observed at

0.5μM and higher of CLK inhibitor. Gene expression regulation, therefore, shows a similar sensitivity to CLK inhibition as seen in AS and CG regulation. A notable exception is a group of genes strongly down-regulated beginning at 5.0μM. These genes may be more resilient to CLK inhibition, or their down-regulation may be a secondary response to strong RNA processing disruption.

Splicing factors were among the genes most affected by AS changes at low doses of T3, indicating that RNA splicing auto-regulation is one of the cellular processes most sensitive to CLK inhibition. Splicing and other RNA processing factors were also involved in CG formation, and their expression was down-regulated in treated cells. One method of splicing factor auto-regulation is the inclusion of a "poison" exon that includes a premature termination codon, and the resulting degradation of the poisoned transcript [60]. AS changes and CG formation may lead to the inclusion of premature termination codons, resulting in reduced expression of RNA processing factors and other genes.

CLK inhibition may also result in cell cycle disruption. Cell cycle progression is linked to RNA splicing, and knock-down of splicing factors can cause cell cycle arrest [61, 62]. Cell cycle related genes were not only differentially spliced, but also participated in CG transcription and were generally down-regulated in treated cells. Therefore, global disruption of splicing through CLK inhibition may interfere with normal cell cycle progression.

High doses of T3 CLK inhibitor may cause pathological cell death. Toll-like receptor ligands released from dead and dying cells may have caused an innate immune response in nearby cells [63], explaining the effects on genes in the toll-like receptor signaling pathway observed in samples treated with high concentrations of T3.

A noticeable similarity in the biological processes affected by CLK inhibition was observed in the analysis of differentially spliced and expressed genes, and CG participants. RNA metabolism (*e.g.* transcription and splicing), cell cycle progression, and protein degradation were among those processes sensitive to loss of CLK activity. Disruption of SR protein activity may cause defects in splicing and transcription [69] (and maybe 3′-end processing) in a common set of genes, which would explain the similarity in affected biological processes.

3.1 Limitations and Future Directions

Alternative splicing analysis was performed using event annotations provided by MISO. These annotations only include a limited set of events derived from expressed sequence tags and gene annotation databases. CLK inhibition may cause splicing defects even in constitutive gene regions and so the MISO annotations may be too restrictive and may have prevented the capture of the full set of splicing changes present in treated cells. Further study into the effects of CLK inhibition on AS would benefit by performing differential splicing analysis on a more comprehensive set of potential AS events, including those which would not undergo differential splicing under normal conditions.

In this thesis, ESE density was shown to correlate with SE and RI splicing response. However, only SE and RI event types were tested and there are likely to be other genomic features predictive of splicing response. Future investigation may be able to predict changes in splicing upon CLK inhibition by inspecting a larger set of features, such as those used in splicing code studies [71], on the full spectrum of AS event types.

Analysis presented in this thesis suggests that SR protein or CFIm phosphorylation may be important for the 3′-end cleavage reaction and CG transcription regulation. However, further experiments are necessary to fully illuminate the role of CLKs in CG formation. One approach might be to use HITS-CLIP assays (cross-linking and immunoprecipitation combined with high-throughput sequencing) to compare RNA processing factor binding profiles in untreated and treated cells. Likewise, immunoprecipitation methods could be used to investigate changes in protein-protein interactions between and with 3′-end processing factors. Further, the proportion of CG transcripts translated into proteins may be tested experimentally. This would shed light on whether CG transcription is primarily a gene expression regulatory mechanism, or it is intended to produce functional proteins. Similar experiments could be performed to fully reveal the mechanism by which CLK inhibition disrupts alternative splicing.

A manuscript of the presented work is in preparation with the intent to submit to a scientific journal.

3.2 Conclusions

This is the first systematic analysis of the transcriptomic consequences of CLK inhibition. Loss of CLK function resulted in the the disruption of RNA splicing, 3′-end processing, and gene expression for genes involved in a common set of biological processes. The dependence of transcript 3′-end cleavage on CLK activity has not been previously reported in the literature. Insights derived from this thesis' will inform future investigations into RNA processing regulation, and the role of CLKs therein.

Bibliography

[1] Pertea M, Salzberg SL (2010) Between a chicken and a grape: estimating the number of human genes. Genome Biol 11: 206. → pages

[2] Harrow J, Frankish A, Gonzalez JM, Tapanari E, Diekhans M, et al. (2012) GENCODE: the reference human genome annotation for The ENCODE Project. Genome research 22: 1760–1774. → pages

[3] Ni JZ, Grate L, Donohue JP, Preston C, Nobida N, et al. (2007) Ultraconserved elements are associated with homeostatic control of splicing regulators by alternative splicing and nonsense-mediated decay. Genes & development 21: 708–718. → pages

[4] Krawczak M, Reiss J, Cooper DN (1992) The mutational spectrum of single base-pair substitutions in mRNA splice junctions of human genes: causes and consequences. Human genetics 90: 41–54. → pages

[5] David CJ, Manley JL (2010) Alternative pre-mRNA splicing regulation in cancer: pathways and programs unhinged. Genes & development 24: 2343–2364. → pages

[6] Yoshida K, Sanada M, Shiraishi Y, Nowak D, Nagata Y, et al. (2011) Frequent pathway mutations of splicing machinery in myelodysplasia. Nature 478: 64–69. → pages

[7] Webb TR, Joyner AS, Potter PM (2013) The development and application of small molecule modulators of SF3b as therapeutic agents for cancer. Drug discovery today 18: 43–49. → pages

[8] Wang Z, Gerstein M, Snyder M (2009) RNA-Seq: a revolutionary tool for transcriptomics. Nature Reviews Genetics 10: 57–63. → pages

[9] Eid J, Fehr A, Gray J, Luong K, Lyle J, et al. (2009) Real-time DNA sequencing from single polymerase molecules. Science 323: 133–138. → pages

[10] Chen M, Manley JL (2009) Mechanisms of alternative splicing regulation: insights from molecular and genomics approaches. Nature Reviews Molecular Cell Biology 10: 741–754. → pages

[11] Shepard PJ, Hertel KJ (2009) The SR protein family. Genome Biol 10: 242. → pages

[12] Shen H, Kan JL, Green MR (2004) Arginine-serine-rich domains bound at splicing enhancers contact the branchpoint to promote prespliceosome assembly. Molecular cell 13: 367–376. → pages

[13] Shen H, Green MR (2004) A pathway of sequential arginine-serine-rich domain-splicing signal interactions during mammalian spliceosome assembly. Molecular cell 16: 363–373. → pages

[14] Cho S, Hoang A, Sinha R, Zhong XY, Fu XD, et al. (2011) Interaction between the RNA binding domains of Ser-Arg splicing factor 1 and U1-70K snRNP protein determines early spliceosome assembly. Proceedings of the National Academy of Sciences 108: 8233–8238. → pages

[15] Ngo JCK, Chakrabarti S, Ding JH, Velazquez-Dones A, Nolen B, et al. (2005) Interplay between SRPK and Clk/Sty kinases in phosphorylation of the splicing factor ASF/SF2 is regulated by a docking motif in ASF/SF2. Molecular cell 20: 77–89. → pages

[16] Long J, Caceres J (2009) The SR protein family of splicing factors: master regulators of gene expression. Biochem J 417: 15–27. → pages

[17] Zhou Z, Fu XD (2013) Regulation of splicing by SR proteins and SR protein-specific kinases. Chromosoma 122: 191–207. → pages

[18] Graveley BR (2004) A protein interaction domain contacts RNA in the prespliceosome. Molecular cell 13: 302–304. → pages

[19] Wang ET, Sandberg R, Luo S, Khrebtukova I, Zhang L, et al. (2008) Alternative isoform regulation in human tissue transcriptomes. Nature 456: 470–476. → pages

[20] Millevoi S, Vagner S (2010) Molecular mechanisms of eukaryotic pre-mRNA 3′ end processing regulation. Nucleic acids research 38: 2757–2774. → pages

[21] Singh G, Kucukural A, Cenik C, Leszyk JD, Shaffer SA, et al. (2012) The cellular EJC interactome reveals higher-order mRNP structure and an EJC-SR protein nexus. Cell 151: 750–764. → pages

[22] Lou H, Neugebauer KM, Gagel RF, Berget SM (1998) Regulation of alternative polyadenylation by U1 snRNPs and SRp20. Molecular and cellular biology 18: 4977–4985. → pages

[23] Iseli C, Stevenson BJ, de Souza SJ, Samaia HB, Camargo AA, et al. (2002) Long-range heterogeneity at the $3'$ ends of human mRNAs. Genome research 12: 1068–1074. → pages

[24] Venkataraman K, Brown KM, Gilmartin GM (2005) Analysis of a noncanonical poly (A) site reveals a tripartite mechanism for vertebrate poly (A) site recognition. Genes & development 19: 1315–1327. → pages

[25] Dettwiler S, Aringhieri C, Cardinale S, Keller W, Barabino SM (2004) Distinct sequence motifs within the 68-kDa subunit of cleavage factor Im mediate RNA binding, protein-protein interactions, and subcellular localization. Journal of Biological Chemistry 279: 35788–35797. → pages

[26] Ryan K (2007) Pre-mRNA $3'$cleavage is reversibly inhibited in vitro by cleavage factor dephosphorylation. RNA biology 4: 26. → pages

[27] Nunes NM, Li W, Tian B, Furger A (2010) A functional human poly (A) site requires only a potent DSE and an A-rich upstream sequence. The EMBO journal 29: 1523–1536. → pages

[28] López-Bigas N, Audit B, Ouzounis C, Parra G, Guigó R (2005) Are splicing mutations the most frequent cause of hereditary disease? FEBS letters 579: 1900–1903. → pages

[29] Krawczak M, Thomas NS, Hundrieser B, Mort M, Wittig M, et al. (2007) Single base-pair substitutions in exon–intron junctions of human genes: nature, distribution, and consequences for mRNA splicing. Human mutation 28: 150–158. → pages

[30] Slaugenhaupt SA, Blumenfeld A, Gill SP, Leyne M, Mull J, et al. (2001) Tissue-Specific Expression of a Splicing Mutation in the IKBKAP Gene Causes Familial Dysautonomia. The American Journal of Human Genetics 68: 598–605. → pages

[31] Nielsen KB, Sørensen S, Cartegni L, Corydon TJ, Doktor TK, et al. (2007) Seemingly Neutral Polymorphic Variants May Confer Immunity to Splicing-Inactivating Mutations: A Synonymous SNP in Exon 5 of MCAD Protects from Deleterious Mutations in a Flanking Exonic Splicing Enhancer. The American Journal of Human Genetics 80: 416–432. → pages

[32] Papaemmanuil E, Cazzola M, Boultwood J, Malcovati L, Vyas P, et al. (2011) Somatic SF3B1 mutation in myelodysplasia with ring sideroblasts. New England Journal of Medicine 365: 1384–1395. → pages

[33] Narla G, DiFeo A, Fernandez Y, Dhanasekaran S, Huang F, et al. (2008) KLF6-SV1 overexpression accelerates human and mouse prostate cancer progression and metastasis. The Journal of clinical investigation 118: 2711–2721. → pages

[34] Ward AJ, Cooper TA (2010) The pathobiology of splicing. The Journal of pathology 220: 152–163. → pages

[35] Ghigna C, Giordano S, Shen H, Benvenuto F, Castiglioni F, et al. (2005) Cell Motility Is Controlled by SF2/ASF through Alternative Splicing of the Ron Protooncogene. Molecular cell 20: 881–890. → pages

[36] Elkon R, Ugalde AP, Agami R (2013) Alternative cleavage and polyadenylation: extent, regulation and function. Nature Reviews Genetics 14: 496–506. → pages

[37] Fedorov O, Huber K, Eisenreich A, Filippakopoulos P, King O, et al. (2011) Specific CLK inhibitors from a novel chemotype for regulation of alternative splicing. Chemistry & biology 18: 67–76. → pages

[38] Chin CS, Alexander DH, Marks P, Klammer AA, Drake J, et al. (2013) Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. Nature methods 10: 563–569. → pages

[39] Quail MA, Smith M, Coupland P, Otto TD, Harris SR, et al. (2012) A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. BMC genomics 13: 341. → pages

[40] Vandenbroucke II, Vandesompele J, De Paepe A, Messiaen L (2001) Quantification of splice variants using real-time PCR. Nucleic Acids Research 29: e68–e68. → pages

[41] Sanger F, Coulson AR (1975) A rapid method for determining sequences in dna by primed synthesis with dna polymerase. Journal of molecular biology 94: 441–448. → pages

[42] Wu TD, Nacu S (2010) Fast and SNP-tolerant detection of complex variants and splicing in short reads. Bioinformatics 26: 873–881. → pages

[43] Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, et al. (2013) Star: ultrafast universal rna-seq aligner. Bioinformatics 29: 15–21. → pages

[44] Wu TD, Watanabe CK (2005) GMAP: a genomic mapping and alignment program for mRNA and EST sequences. Bioinformatics 21: 1859–1875. → pages

[45] A hands on tutorial of three aligners: BLAT, BLASR, and GMAP. https://github.com/PacificBiosciences/cDNA_primer/wiki/A-hands-on-tutorial-of-three-aligners%3A-BLAT%2C-BLASR%2C-and-GMAP. Accessed: 2014-07-15. → pages

[46] Katz Y, Wang ET, Airoldi EM, Burge CB (2010) Analysis and design of RNA sequencing experiments for identifying isoform regulation. Nature methods 7: 1009–1015. → pages

[47] Alamancos GP, Agirre E, Eyras E (2013) Methods to study splicing from high-throughput RNA Sequencing data. arXiv preprint arXiv:13045952 . → pages

[48] Anders S, Huber W (2010) Differential expression analysis for sequence count data. Genome biol 11: R106. → pages

[49] Robinson MD, McCarthy DJ, Smyth GK (2010) edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 26: 139–140. → pages

[50] Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nature methods 5: 621–628. → pages

[51] Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, et al. (2010) Transcript assembly and quantification by RNA-seq reveals unannotated transcripts and isoform switching during cell differentiation. Nature biotechnology 28: 511–515. → pages

[52] Roberts A, Trapnell C, Donaghey J, Rinn JL, Pachter L (2011) Improving RNA-Seq expression estimates by correcting for fragment bias. Genome biology 12: R22. → pages

[53] Muraki M, Ohkawara B, Hosoya T, Onogi H, Koizumi J, et al. (2004) Manipulation of alternative splicing by a newly developed inhibitor of Clks. Journal of Biological Chemistry 279: 24246–24254. → pages

[54] Barash Y, Calarco JA, Gao W, Pan Q, Wang X, et al. (2010) Deciphering the splicing code. Nature 465: 53–59. → pages

[55] McPherson A, Hormozdiari F, Zayed A, Giuliany R, Ha G, et al. (2011) deFuse: an algorithm for gene fusion discovery in tumor RNA-seq data. PLoS computational biology 7: e1001138. → pages

[56] Wu G, Feng X, Stein L (2010) A human functional protein interaction network and its application to cancer data analysis. Genome Biol 11: R53. → pages

[57] Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, et al. (2000) Gene ontology: tool for the unification of biology. Nature genetics 25: 25–29. → pages

[58] Cartegni L, Wang J, Zhu Z, Zhang MQ, Krainer AR (2003) ESEfinder: a web resource to identify exonic splicing enhancers. Nucleic acids research 31: 3568–3571. → pages

[59] Merico D, Isserlin R, Stueker O, Emili A, Bader GD (2010) Enrichment map: a network-based method for gene-set enrichment visualization and interpretation. PloS one 5: e13984. → pages

[60] Anko ML, Muller-McNicoll M, Brandl H, Curk T, Gorup C, et al. (2012) The RNA-binding landscapes of two SR proteins reveal unique functions and binding to diverse RNA classes. Genome Biol 13: R17. → pages

[61] Blencowe BJ (2003) Splicing regulation: the cell cycle connection. Current biology 13: R149–R151. → pages

[62] Kurokawa K, Akaike Y, Masuda K, Kuwano Y, Nishida K, et al. (2013) Downregulation of serine/arginine-rich splicing factor 3 induces G1 cell cycle arrest and apoptosis in colon cancer cells. Oncogene 33: 1407–1417. → pages

[63] Kawai T, Akira S (2010) The role of pattern-recognition receptors in innate immunity: update on Toll-like receptors. Nature immunology 11: 373–384. → pages

[64] Langfelder P, Horvath S (2008) WGCNA: an R package for weighted correlation network analysis. BMC bioinformatics 9: 559. → pages

[65] Langfelder P, Horvath S. Wgcna package faq. http://labs.genetics.ucla.edu/horvath/CoexpressionNetwork/Rpackages/WGCNA/faq.html. Accessed: 2014-11-12. → pages

[66] Thorvaldsdóttir H, Robinson JT, Mesirov JP (2012) Integrative genomics viewer (igv): high-performance genomics data visualization and exploration. Briefings in bioinformatics : bbs017. → pages

[67] Prakash T, Sharma VK, Adati N, Ozawa R, Kumar N, et al. (2010) Expression of conjoined genes: another mechanism for gene regulation in eukaryotes. PloS one 5: e13284. → pages

[68] Shin C, Manley JL (2002) The SR protein SRp38 represses splicing in M phase cells. Cell 111: 407–417. → pages

[69] Lin S, Coutinho-Mansfield G, Wang D, Pandit S, Fu XD (2008) The splicing factor SC35 has an active role in transcriptional elongation. Nature structural & molecular biology 15: 819–826. → pages

[70] Greger L, Su J, Rung J, Ferreira PG, Lappalainen T, et al. (2014) Tandem RNA Chimeras Contribute to Transcriptome Diversity in Human Population and Are Associated with Intronic Genetic Variants. PloS one 9: e104567. → pages

[71] Leung MK, Xiong HY, Lee LJ, Frey BJ (2014) Deep learning of the tissue-regulated splicing code. Bioinformatics 30: i121–i129. → pages

Appendix A

Supporting Materials

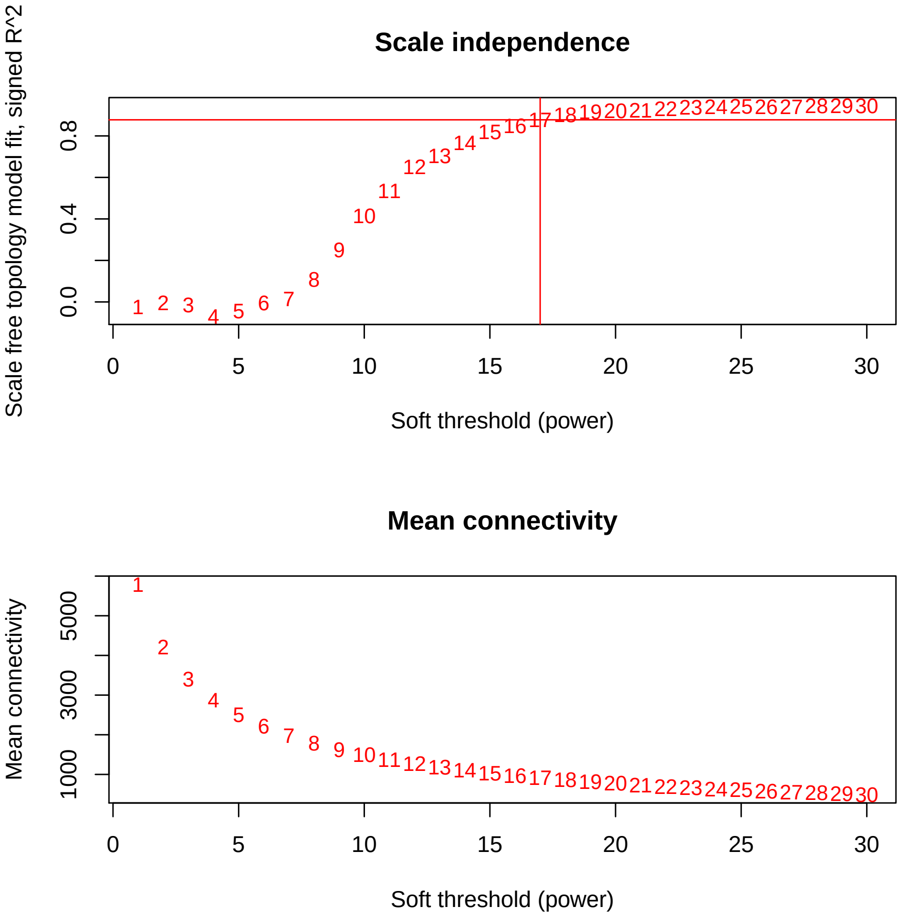

Figure A.1: Soft threshold vs. scale independence and vs. mean connectivity for HCT116 unstranded RNA-Seq AS PSI WGCNA clustering.

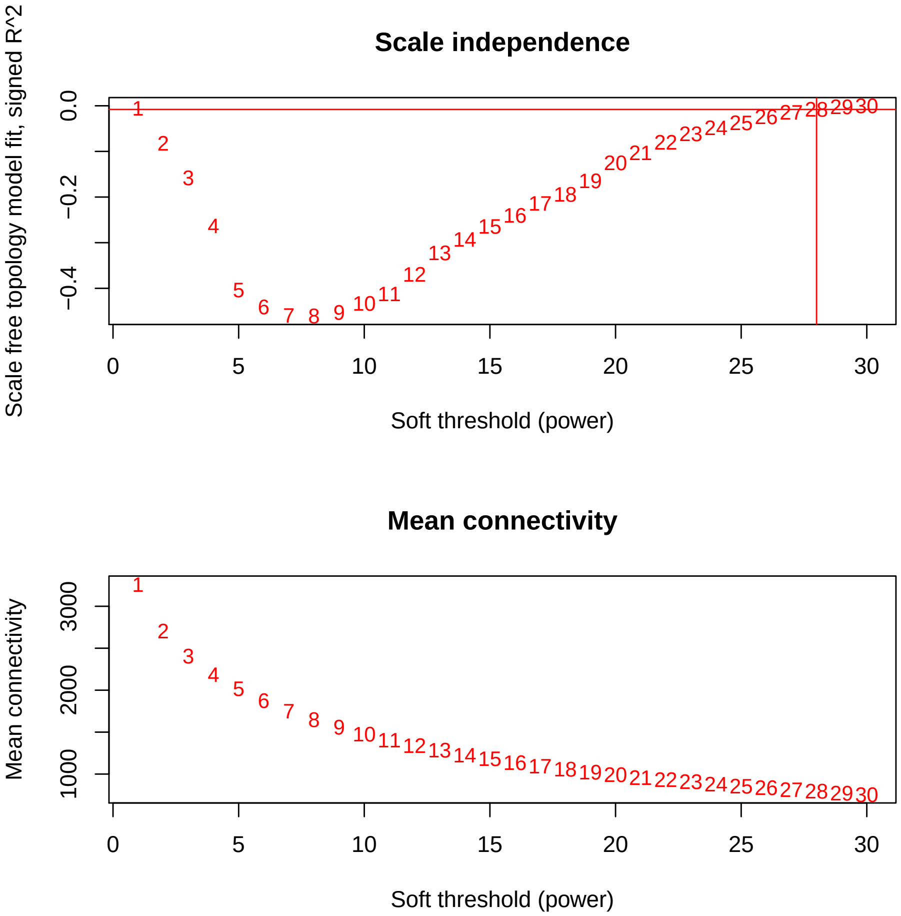

Figure A.2: Soft threshold vs. scale independence and vs. mean connectivity for HCT116 stranded RNA-Seq AS PSI WGCNA clustering.

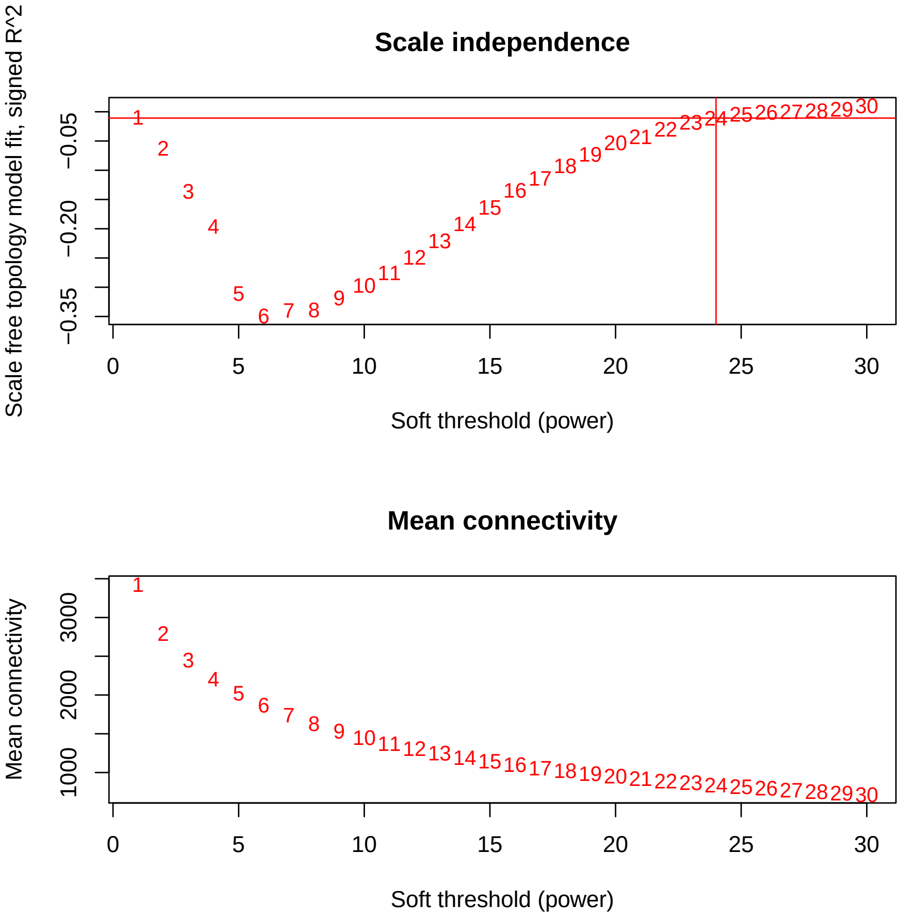

Figure A.3: Soft threshold vs. scale independence and vs. mean connectivity for hTERT stranded RNA-Seq AS PSI WGCNA clustering.

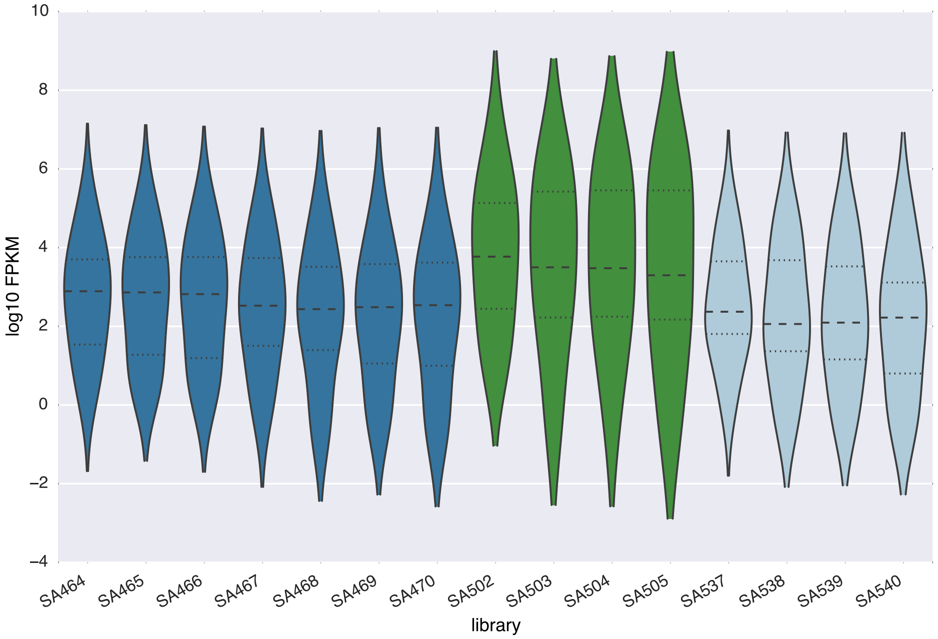

Figure A.4: Violin plots of $\log_{10} FPKM$ values for upstream gene partners of hTERT exclusive conjoined genes. FPKM values are plotted for both HCT116 RNA-Seq datasets and the hTERT dataset. SA464-470 are the HCT116 samples used for unstranded RNA-Seq (dark blue), SA537-540 are HCT116 samples used for stranded RNA-Seq (light blue), and SA502-505 are hTERT samples (green). Violin plots for each dataset are ordered by increasing T3 concentration.

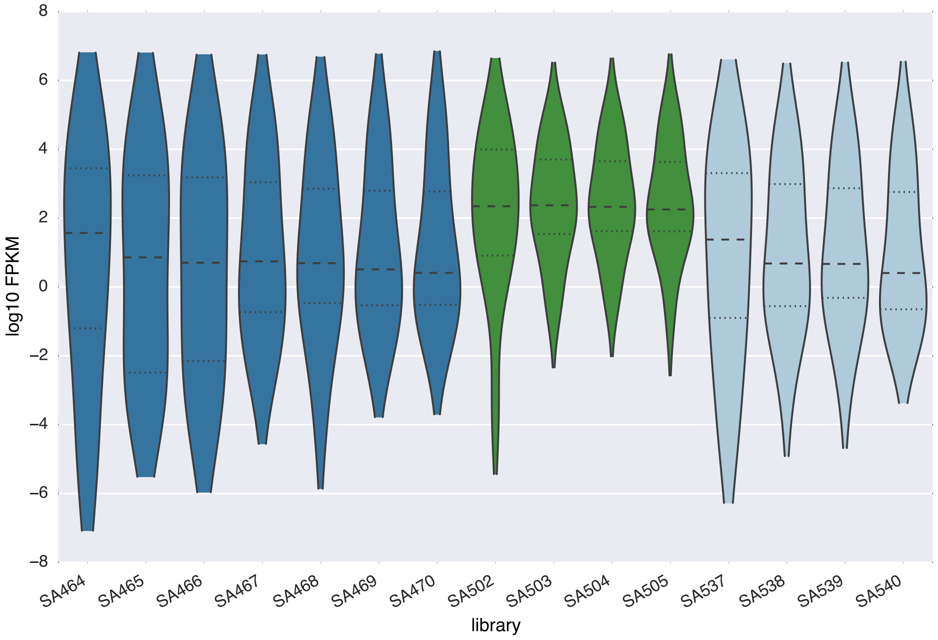

Figure A.5: Violin plots of $\log_{10} FPKM$ values for downstream gene partners of hTERT exclusive conjoined genes. FPKM values are plotted for both HCT116 RNA-Seq datasets and the hTERT dataset. SA464-470 are the HCT116 samples used for unstranded RNA-Seq (dark blue), SA537-540 are HCT116 samples used for stranded RNA-Seq (light blue), and SA502-505 are hTERT samples (green). Violin plots for each dataset are ordered by increasing T3 concentration.

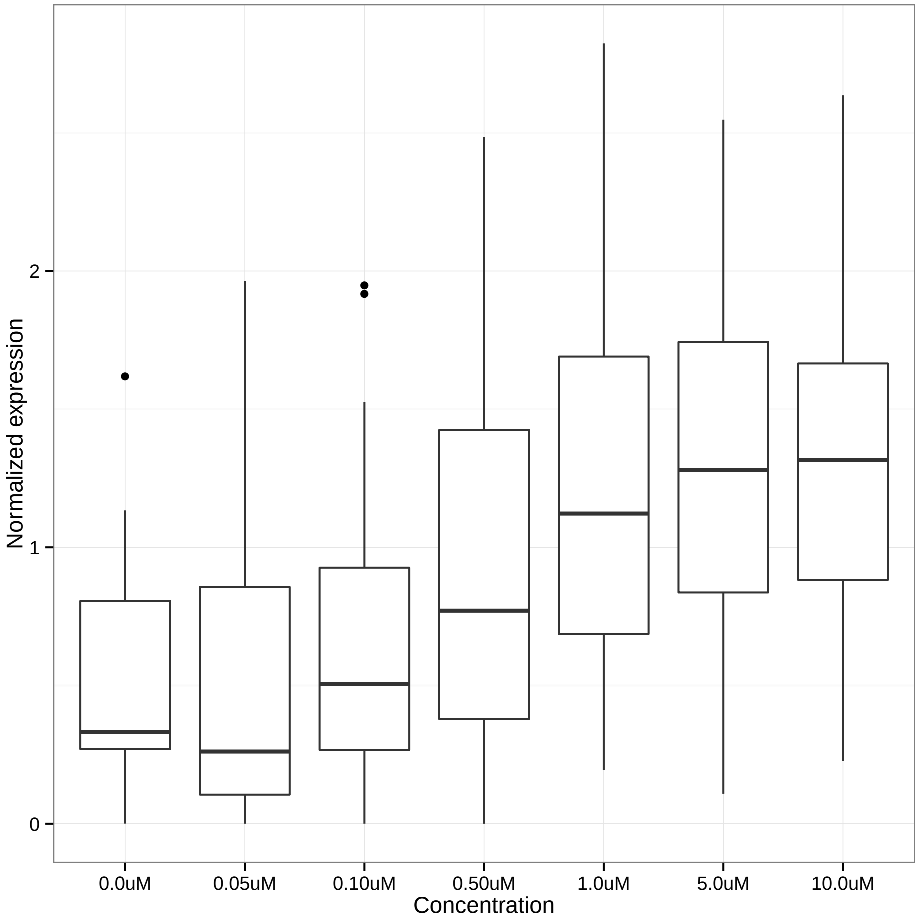

Figure A.6: Normalized conjoined gene expression boxplots across T3 concentrations for HCT116 replicate 1 dataset. Conjoined gene expression has been normalized to ACTB expression. This dataset is generated from the same samples used to generate the HCT116 unstranded RNA-Seq dataset.

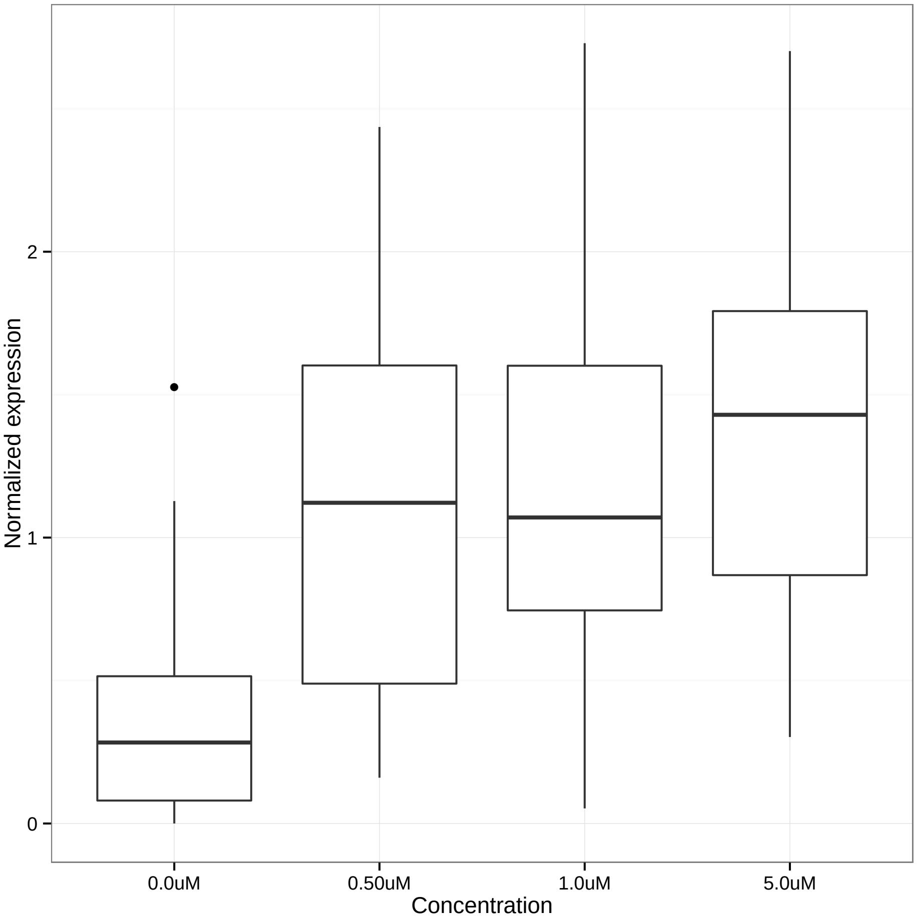

Figure A.7: Normalized conjoined gene expression boxplots across T3 concentrations for HCT116 replicate 2 dataset. Conjoined gene expression has been normalized to ACTB expression. This dataset is generated from the same samples used to generate the HCT116 stranded RNA-Seq dataset.

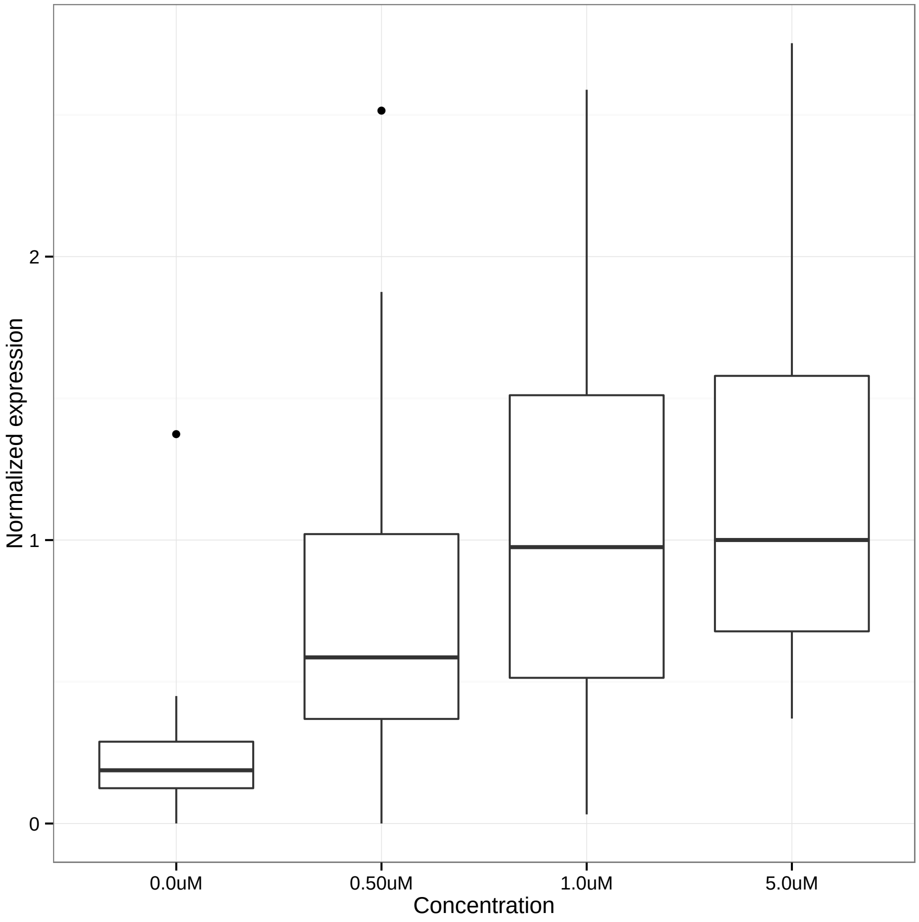

Figure A.8: Normalized conjoined gene expression boxplots across T3 concentrations for hTERT dataset. Conjoined gene expression has been normalized to ACTB expression. This dataset is generated from the same samples used to generate the hTERT stranded RNA-Seq dataset.

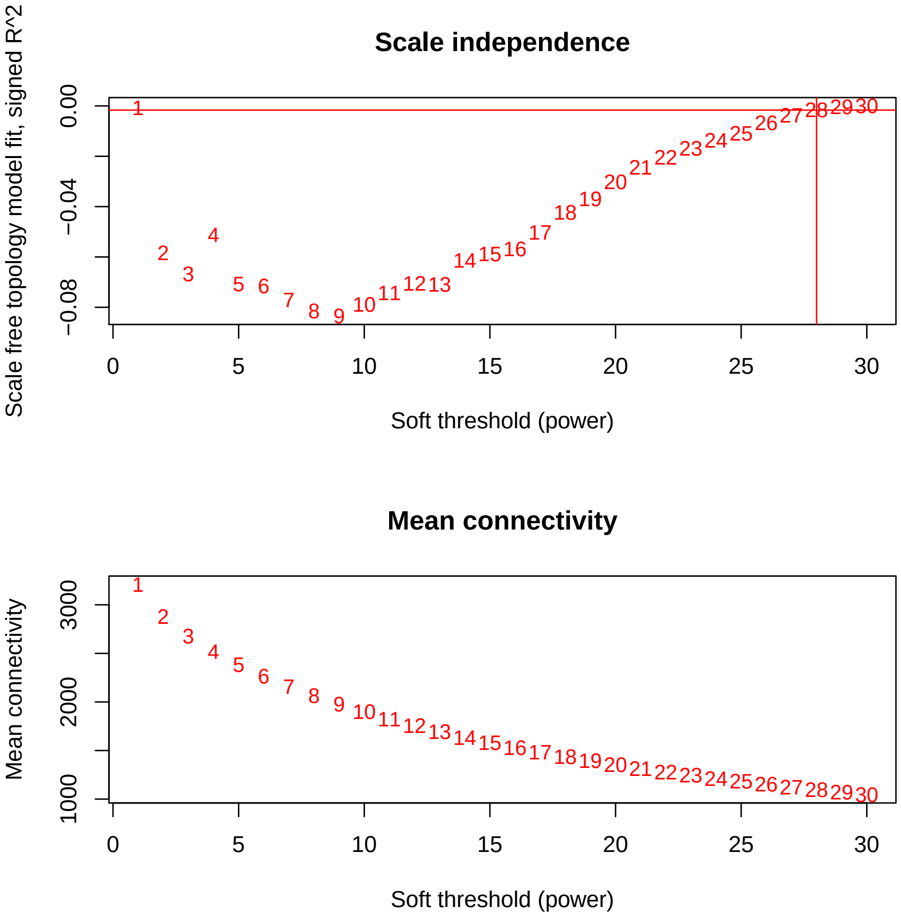

Figure A.9: Soft threshold vs. scale independence and vs. mean connectivity for HCT116 unstranded RNA-Seq FPKM WGCNA clustering.

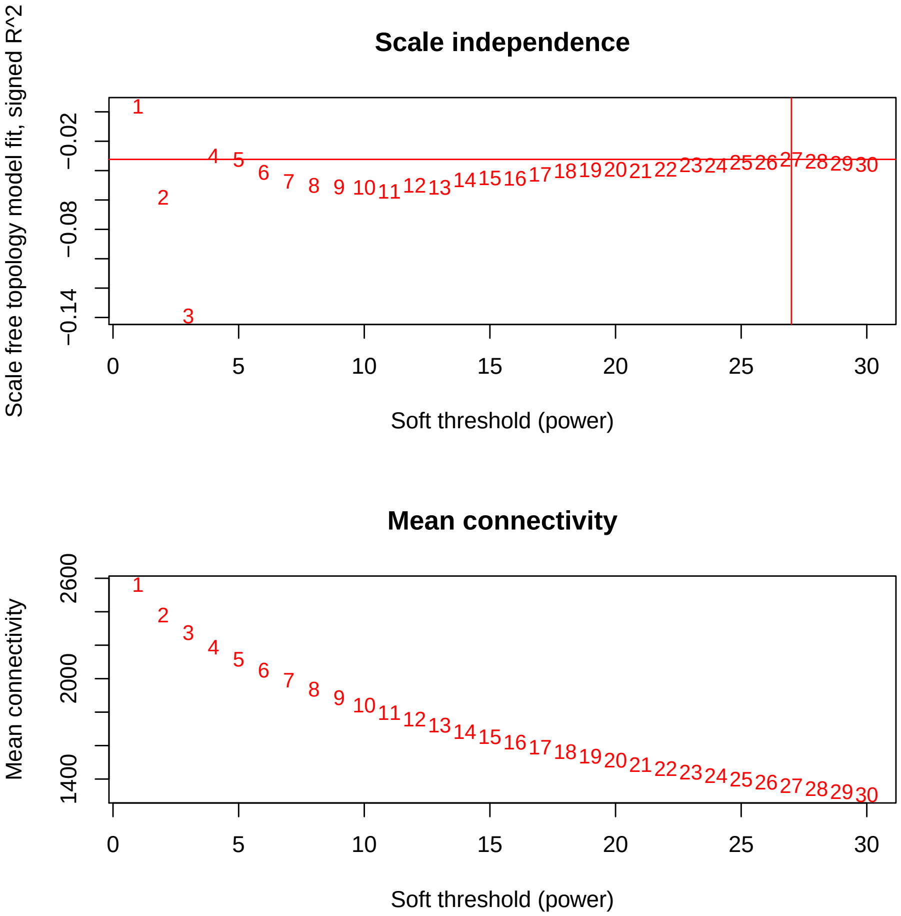

Figure A.10: Soft threshold vs. scale independence and vs. mean connectivity for HCT116 stranded RNA-Seq FPKM WGCNA clustering.

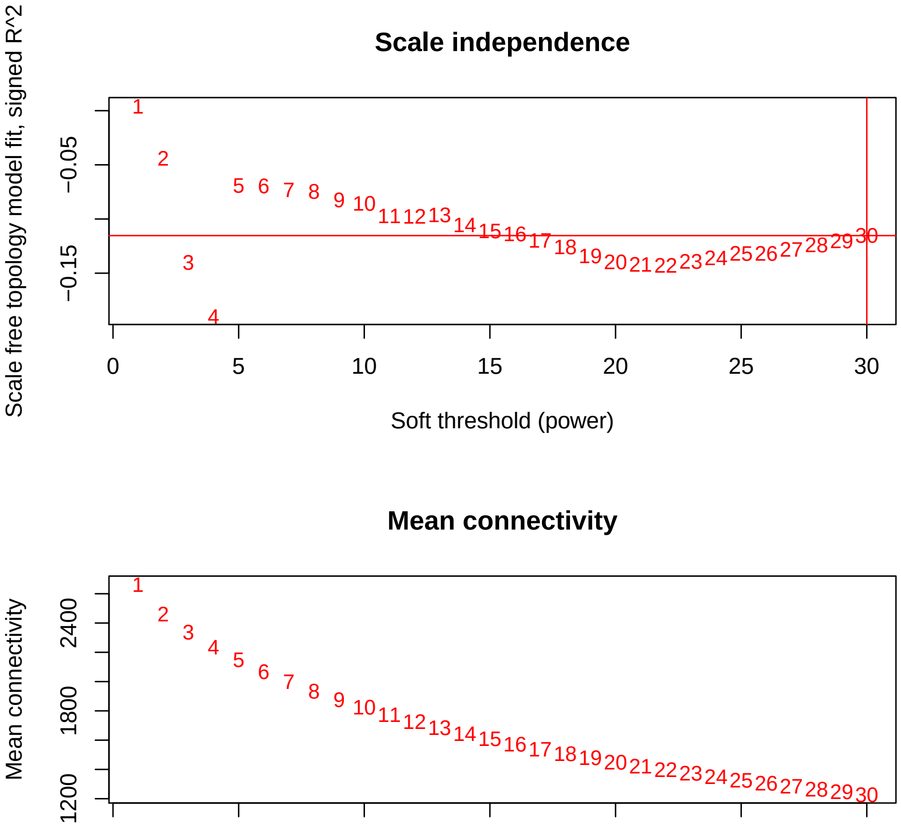

Figure A.11: Soft threshold vs. scale independence and vs. mean connectivity for hTERT stranded RNA-Seq FPKM WGCNA clustering.