ALTERNATIVE MRNA SPLICING REDEFINES THE LANDSCAPE OF COMMONLY DYSREGULATED GENES ACROSS THE ACUTE MYELOID LEUKEMIA PATIENT POPULATION

Osvaldo D. Rivera

A DISSERTATION

in

Cell and Molecular Biology

Presented to the Faculties of the University of Pennsylvania

in

Partial Fulfillment of the Requirements for the

Degree of Doctor of Philosophy

2021

Supervisor of Dissertation

Kristen W. Lynch, Ph.D.

Professor and Chair

Department of Biochemistry & Biophysics

Supervisor of Dissertation

Yoseph Barash, Ph.D.

Associate Professor of Genetics

Department of Genetics

Graduate Group Chairperson

Daniel S. Kessler, Ph.D., Associate Professor of Cell and Molecular Biology

Dissertation Committee

Martin P. Carroll, MD, Associate Professor of Medicine

Peter Choi, Ph.D., Assistant Professor of Pathology and Laboratory Medicine.

Robert Babak Faryabi, Ph.D., Assistant Professor of Pathology and Laboratory Medicine.

Kathy Fange Liu, Ph.D., Assistant Professor of Biochemistry and Biophysics

ACKNOWLEDGEMENTS

Special thanks to my advisors Dr. Kristen W. Lynch and Dr. Yoseph Barash. Their continuous support and expert guidance have been an invaluable asset to my development of technical skill and critical thinking. They both taught me how to harness my potential and to pick myself up from the constant challenges of research and of life. Thanks to the rest of my Thesis Committee for their expert insight and support during key milestones.

Special thanks to Dr. Martin P. Carroll, who supported me when I was doubtful of my progress, and always kept in touch on my whereabouts and actively contributes to the development my career. I want to also acknowledge Martin as Medical Director of the Stem Cell and Xenograft Core of the Perelman School of Medicine and the Hematologic Malignancies Translational Center of Excellence of the Abramson Cancer Center, which collected and stored the AML patient samples used for this study.

Special thanks to Dr. Sara Cherry for her significant intellectual contribution to my project as well as her collaborative work with Dr. David Schultz, both directors of the Penn High-Throughput Screening Core (HTSC). Thanks to the HTSC core staff for thawing and culturing the AML patient blast cells with high confirmed viability and purity, as well as ordering and providing the sequencing of the mRNA extracts used for this study.

Special thanks to Lynch Lab member Michael Mallory for performing radiolabeled RT-PCR and Western blots in the patient AML samples and to Lynch Lab member Dr. Rakesh Chatrikhi for engineering the cell lines used in my functional studies of DHX34 poison exon inclusion. Also, special thanks to Dr. Mathieu Quesnel-Vallieres for helping process the large number of BeatAML RNA-seq samples. Thanks to all of the Lynch Lab and Barash Lab colleagues who facilitated my learning of new skills. It was an honor and a privilege to work alongside all of you.

Special thanks to the National Science Foundation (NSF) Graduate Research Fellowship Program (GRFP) and the University of Pennsylvania for their financial and educational support during my research training. Special thanks to professors Dr. Daniel S. Kessler, Dr. Sandra W. Ryeom, and Dr. Craig H. Bassing as well as my rotation mentors at the University of Pennsylvania for leading an exceptional graduate program in Cell and Molecular Biology and Cancer Biology. It was a privilege for me to have the support of the graduate group offices throughout the duration of my research training.

Special thanks to Dr. Arnaldo J. Diaz and the Office for Research & Diversity Training for their support throughout a large portion of my scientific training. Arnaldo's mentoring programs were an invaluable asset to the development of my career.

Lastly, special thanks to my parents, my loving family, and my close friends for their unconditional support since the beginning of my journey as a Research Scientist. Thanks for allowing me to share my struggles and challenges and helping me stay grounded with my work. I did everything at the right time thanks to all of you.

I dedicate this dissertation to all of you, my supporters, who always motivated me to pursue higher levels of knowledge and become the first in my family to complete a Ph.D. and develop into a successful Scientist

ABSTRACT

ALTERNATIVE MRNA SPLICING REDEFINES THE LANDSCAPE OF COMMONLY DYSREGULATED GENES ACROSS THE ACUTE MYELOID LEUKEMIA PATIENT POPULATION

Osvaldo D. Rivera

Kristen W. Lynch & Yoseph Barash

Most genes associated with acute myeloid leukemia (AML) are mutated in less than 10% of patients, suggesting that alternative mechanisms of gene disruption contribute to this disease. Here, I investigate pre-mRNA splicing events with significant variation and striking coregulation across distinct AML cohorts. I find that most splicing events are expected to alter the expression of a subset of AML-associated genes independent of known somatic mutations. In particular, I highlight that aberrant splicing triples the number of patients with reduced functional EZH2 protein compared with that predicted by somatic mutation alone. In addition, I unexpectedly find that transcripts encoding the nonsense-mediated RNA decay factor DHX34 exhibit widespread alternative splicing in sporadic AML, resulting in a premature stop codon that phenocopies the loss-of-function germline mutations observed in familial AML. The identification of *DHX34* splicing event that functionally downregulates the nonsense-mediated mRNA degradation (NMD) pathway motivated a query of splicing variations in an additional set of related NMD factors. Although no particular study has highlighted genetic mutations in the queried NMD factors, I find significant variation at the level of mRNA splicing that is expected to have further deleterious effects across AML patients. Together, these results demonstrate that classical mutation analysis underestimates the burden of functional gene disruption in AML and highlight the importance of assessing the contribution of alternative splicing to gene dysregulation in human disease.

Contents

List of Illustrations

CHAPTER I

CHAPTER II

CHAPTER III

CHAPTER IV

List of Tables

CHAPTER II

PREFACE

Acute myeloid leukemia (AML) is an aggressive hematologic cancer in which malignant myeloid precursor cells impair hematopoiesis and induce bone marrow failure. The molecular heterogeneity of AML patients has been recognized for over 35 years now. Most genes associated with AML are mutated in less than 10% of patients, suggesting that alternative mechanisms of gene disruption may contribute to this disease. Although a significant body of work exists detailing the pathology and molecular basis of distinct leukemias, very few studies exist that characterize dysregulation at the level of RNA processing. The maturation of messenger RNA (mRNA) molecules requires distinct process and are largely coordinated through alternative splicing patterns produced by the spliceosome machinery. About 95% of transcribed multi exon genes are spliced in more than one way to give rise to multiple mature transcript varieties from a single genetic locus. This in turn generates complex transcriptomic and proteomic diversity that has been overlooked by most studies of AML and cancer. Furthermore, it wasn't until recently that technologies facilitated the high-throughput quantification of alternative mRNA splicing. There is a growing body of evidence that underscores the importance of dysregulating alternative mRNA splicing in supporting an oncogenic molecular profile and generating splice variants that phenocopy the effects of genetic mutations. The following dissertation documents analyses that shed light on patterns of alternative splicing within genes associated with AML and demonstrate that alternative splicing provides an additional mechanism by which gene function is disrupted in AML. These results demonstrate that classical mutation analysis underestimates molecular burden of functional gene disruption in AML and highlight the importance of assessing the contribution of alternative splicing to gene dysregulation in human disease.

CHAPTER I

Introduction

RNA Biology

Pre-mRNA Transcription and Splicing. RNA transcription is an early essential the step in the process of successful gene expression. In eukaryotes, transcription is executed by one of three RNA polymerases. Specifically, protein-coding genes are transcribed by RNA polymerase II (RNAPII) into premature-mRNA (pre-mRNA) in a strand-specific manner. RNAPII transcription initiates at millions of positions in the mammalian genome from core promoters that have an array of closely located transcription-start sites (TSSs) with different rates of initiation [1]. RNAPII moves along genes that are up to millions of base pairs in length and generates pre-RNA as it transcribes the gene region. The transcription machinery and the nascent pre-mRNA interact with various small-nuclear ribonucleoproteins (snRNPs) and RNA-binding proteins (RBPs) to secure the maturation of pre-mRNA. The C-terminal domain tail of RNAPII regulates the activity of RNA-protein complexes involved in RNA processing and maturation steps such as 5'-end capping, splicing, 3'-end processing, editing, folding, nuclear export and decay of mRNA [2]. Splicing of mRNA precursors is therefore a co-transcriptional event required for the maturation of almost all human mRNAs and is a key step in the regulation of expression of many genes.

During the process of pre-mRNA splicing, a collection of snRNPs that make up the spliceosome bind to conserved RNA sequence features known as splice-sites and define intron/exon boundaries across gene transcripts. Specifically, the major spliceosome is a megadalton complex composed of multiple distinct snRNPs (U1, U2, U4, U5, U6) [3]. The RNA splicing mechanics of more than 99% of protein-coding transcripts are essentially regulated by the major spliceosome while fewer than 1,000 introns (i.e., ~0.3%) are removed by the minor spliceosome, which uses distinct snRNPs (U11, U12, U4atac, and

U6atac) but shares U5 and most proteins with the major spliceosome [4]. The process of pre-mRNA splicing involves the stepwise interaction between the spliceosome and a set of sequence motifs known as the core splicing signals, namely the 5′ splice site (5′ss), the 3′ splice site (3′ss), the branch point sequence (BPS), and the polypyrimidine tract (Py-tract). The recruitment and activity of the spliceosome constitute the usage of both core splicing signals and secondary *cis*-acting splicing RNA elements (SREs) followed by a cascade of molecular remodeling events that can involve more than 100 others auxiliary RBPs known as "splicing factors" [5]. Additionally, transcription elongation rates driven by gene architecture and chromatin features can in turn influence splice site identification by the spliceosome and modulate splicing patterns [6].

The earliest steps of the metazoan spliceosome assembly are built around the exon junction boundaries in a process called "exon definition". The inclusion of most exons is largely under control by the inherent strength or weakness of the flanking core splicing signals as well as combinatorial binding of auxiliary splicing factors to distinct SREs. [7; 8; 9; 10; 11] (**Figure 1.1A**). The context and positional relationship between core splicing signals and SREs are crucial for the exon-definition process, particularly across gene transcripts with complex architecture that undergo tightly regulated alternative splicing events. Accordingly, sequences comprising this second class of *cis*-acting RNA elements can be subdivided into splicing '*silencers*' and '*enhancers*' that often overlap across exon and intron regions. Depending on the location and function of the particular SRE, it is categorized as an exonic (ESE/ESI) or intronic (ISE/ISS) splicing enhancer or silencer (**Figure 1.1A**). Most studies highlight that the 200–300 nucleotides adjacent to the observed splice sites, harbor the most relevant sequence features for the modulation of

alternative splicing [12; 13]. However, in some cases, there is evidence that distal intronic SREs (> 500 nt. from the exon being spliced) are able to facilitate splicing regulation [14].

After binding to the pre-mRNA has occurred, the spliceosome transitions into subcomplexes that direct the RNA splicing reaction via ATP-dependent conformational rearrangements and molecular interactions with the coupled auxiliary splicing factors [3]. Initially, the U1 snRNP binds to the 5'ss and the U2 auxiliary factor coupled with U2 snRNP binds to the Py-tract/3'ss to form the *pre-spliceosome* and positions the ends of the intron into a catalytic center. Subsequently, the U4/U6.U5 tri-snRNPs activate the pre-spliceosome and reconfigure the interactions between the snRNPs and pre-mRNA splice sites. The splicing reaction itself is a two-step transesterification reaction catalyzed by U6/U2 snRNA complex that resembles a self-splicing ribozyme (**Figure 1.2A**). In the first step (step I), the 2′-hydroxyl of the BPS adenosine carries out a nucleophilic attack on the phosphodiester bond of the 5′ss guanosine, yielding a 5′ exon with a free 3′-hydroxyl and a branched intron lariat that is attached to the 3′ exon. In the second step (step II), the 3′-hydroxyl of the 5′ exon attacks the first nucleotide downstream of the 3′ss guanosine, thereby splicing the 5′ss to the 3'-exon and releasing the excised intron lariat

Over 95% of human multi-exon genes can express functionally different RNA transcript isoforms by alternatively including or excluding particular exons, with at least 80% of these changes altering the protein-coding potential of said transcript [15; 11]. Thus, alternative pre-mRNA splicing enables vast proteomic diversity from a limited number of genes. Furthermore, there is a growing body of evidence that coordinated alternative splicing networks regulate important physiological functions in different developmental processes in humans, including tissue and organ development [16]

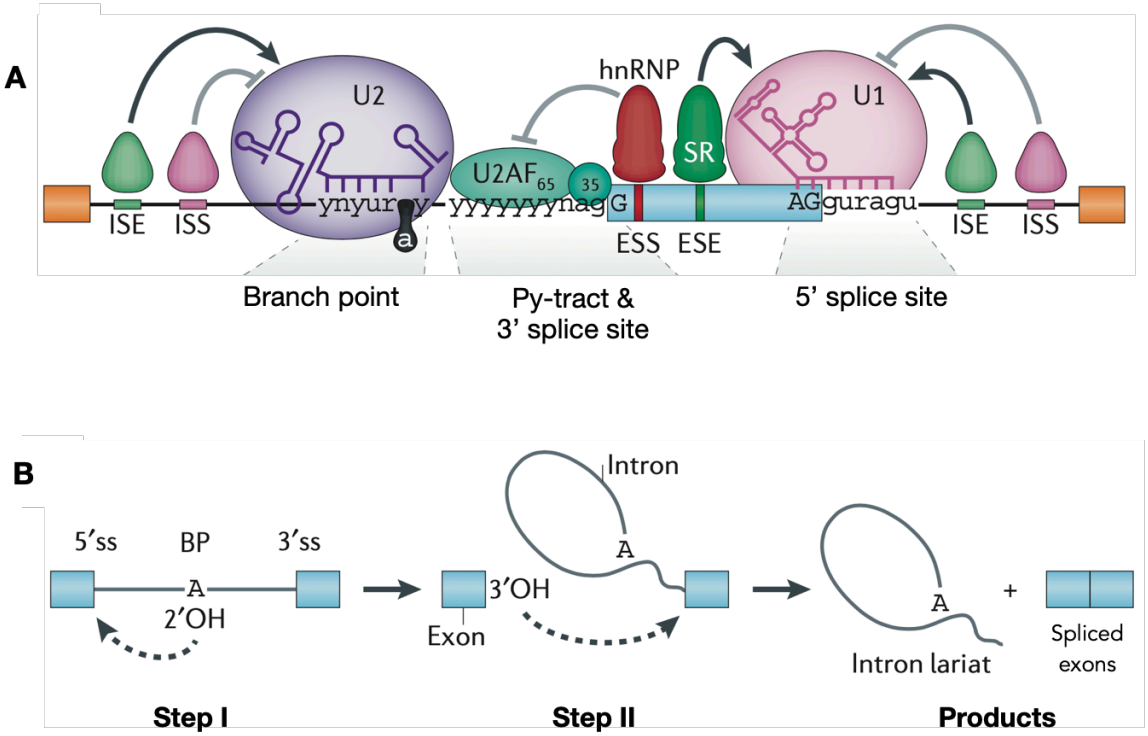
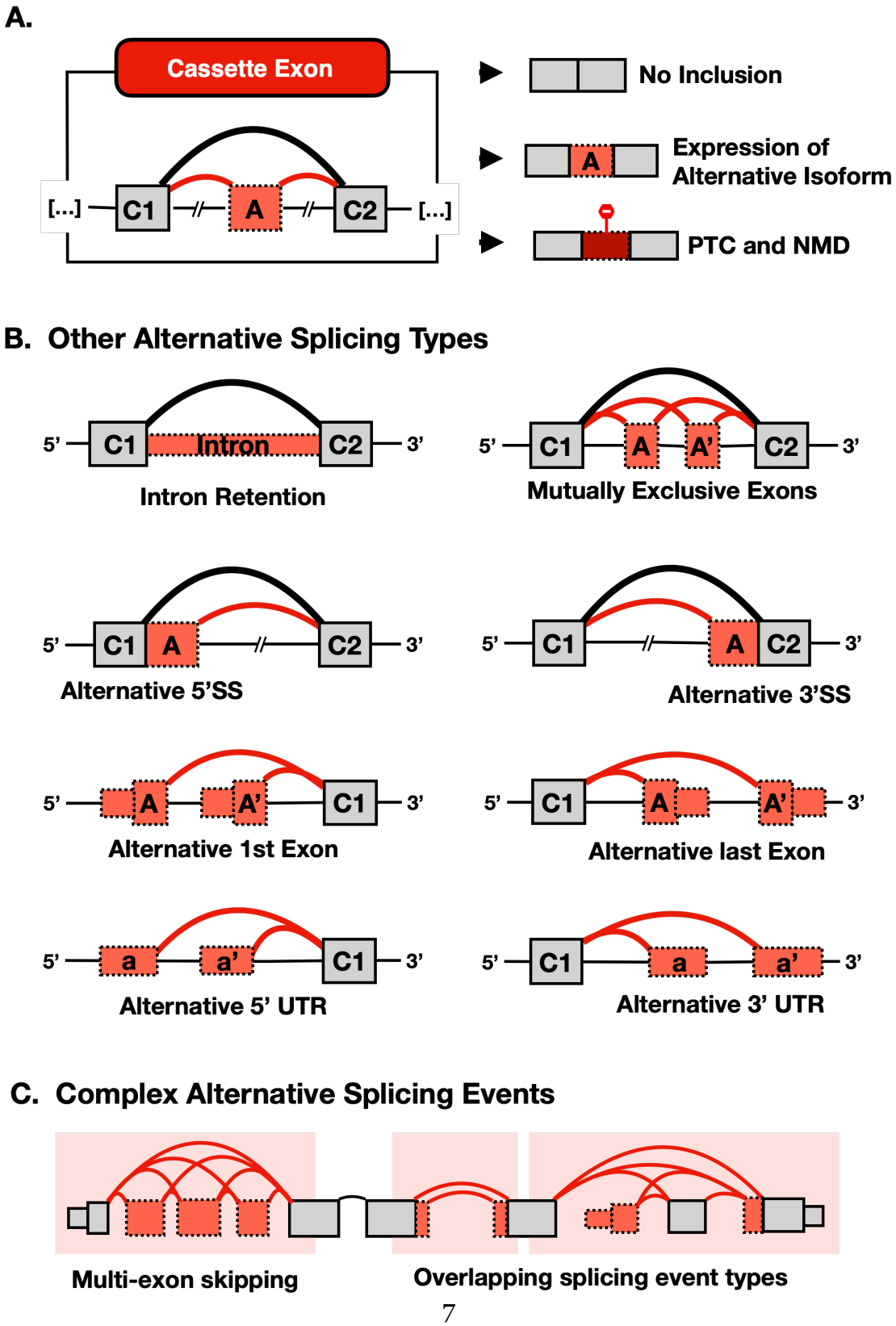


Figure 1.1 – RNA Splicing Reaction, Core Splicing Signals and Alternative splicing types. (**A**) RNA splicing reaction depicted as a sequential transesterification reaction initiated by nucleophilic attack of the 5' splice site by the branch point adenosine in the intron being spliced out. This is followed by a second attack from the 3' hydroxyl group of the 5'ss to the 3' ss that results in the formation of an intron lariat plus the spliced product. (**B**) Spliceosome binding to core splicing signals (BPS, Py-tract, 3'ss and 5'ss) and *cis*-acting SREs – namely exonic and intronic splicing enhancers (ESE/ISE) and silencers (ESS/ISS) sequence features of the pre-mRNA. (Adapted from Mariana & Swanson 2016)

A Code for Alternative Splicing. The cascade of molecular rearrangements that make up the assembly and sub-complex transitions of the spliceosome as well as the recruitment of auxiliary splice factors represent potential points of regulation for the alternative inclusion of particular regions of the pre-mRNA sequence. In alternative splicing events, recognition and joining of a 5'ss-3'ss pair by the spliceosome are in competition with at least one other 5'ss-3'ss pair. Common outcomes of alternative splicing include skipping or inclusion of a cassette exon, alternative 5′ or 3′ splice site choice lengthening or shortening an exon, mutually exclusive exons, alternative first or last exon, and intron retention (**Figure 1.2A-B**). In practice, splicing events may involve several splice site choices yielding complex patterns of alternative splicing (**Figure 1.2C**). Indeed, complex splicing events are common in diverse metazoans, making up at least a third of observed splicing events in human and mouse [17]. Complex splicing events are less likely to have a 'dominant' splice junction (inclusion > 60%). An analysis of datasets across different tissues and developmental stages revealed that complex splicing events have more evenly distributed inclusion levels with no dominant junction as well as an enrichment for regulated splicing of a third junction (inclusion > 10%), suggesting regulation and possible functionality of multiple isoforms [17].

Figure 1.2 – Effects of Alternative Splicing Types (**A**) Cassette Exons differentially includes one particular exon. The binary inclusion on the alternative exon can promote the expression of an alternative protein isoform or insert a premature termination codon (PTC) and tag the transcript for nonsense mediate RNA decay (NMD). C1 and C2 are common nomenclature for the constitutive exons involved in a binary splicing choice while 'A' depicts the alternative sequence. (**B**) Other Common binary alternative splicing event types include intron retention (IR), Mutually exclusive exon inclusion, alternative 5'ss and 3'ss, alternative first and last exon, alternative 5'UTR and 3'UTR (**C**) Complex alternative splicing events have more than two junctions that can be differentially spliced together. Examples include multi-exon skipping events as well as distinct splicing event types that have overlapping splicing junction usage.


A.
Cassette Exon
[...]
C1
A
C2
[...]
No Inclusion
A
Expression of Alternative Isoform
PTC and NMD
B. Other Alternative Splicing Types
5'
C1
Intron
C2
3'
Intron Retention
5'
C1
A
A'
C2
3'
Mutually Exclusive Exons
5'
C1
A
C2
3'
Alternative 5'SS
5'
C1
A
C2
3'
Alternative 3'SS
5'
A
A'
C1
3'
Alternative 1st Exon
5'
C1
A
A'
3'
Alternative last Exon
5'
a
a'
C1
3'
Alternative 5' UTR
5'
C1
a
a'
3'
Alternative 3' UTR
C. Complex Alternative Splicing Events
Multi-exon skipping
Overlapping splicing event types

Combinatorial knockdown of RBPs reveal differences in the recognition of splice sites by selectively influencing the recruitment of the spliceosome [18]. These phenomena underscore how the spliceosome has inherently weak affinity to the RNA substrate, and that additional interactions from multiple auxiliary splicing factors are required to maintain the integrity and fidelity of the pre-mRNA splicing process. Most regulatory contexts of alternative splicing events arise from auxiliary RBPs competitively binding to overlapping sets of *cis*-acting SREs along the transcript. RBPs bind SRE motifs discriminately but also, with a range of redundancy, allowing transcripts to ensure the ability to bind the necessary RBPs required to adhere to the coding standards and global cell regulatory constraints. Gene transcripts are regulated by a discrete subset of RBPs, and each RBP typically regulates several other hundred splicing events, thereby installing alternative splicing networks that allow cells to readily respond to molecular stimuli [10; 11; 19; 20]. Therefore, even modest changes in the abundance or activity of individual RBPs can result in altered splicing patterns involving hundreds of different transcripts. Ultimately, the decision of which splice sites are used to produce a particular splicing pattern is determined by the complex interplay of a pool of *trans*-acting RBPs that competitively bind along exons and introns in a combinatorial fashion thereby influencing the spliceosome through a variety of molecular mechanisms and cellular pathways [21]

In humans, classically identified families of *trans*-acting splicing factors and regulatory RBPs can be divided into three major classes: Serine/arginine-rich (SR) proteins, canonical heterogeneous nuclear ribonucleoproteins (hnRNPs), and other hnRNP-like proteins [22]. SR proteins are defined by containing at least one RRM-type RNA-binding domain and a serine–arginine (SR)-rich domain that facilitates homotypic interactions with core components of the spliceosome as well as other SR proteins [23;11].

Although not always the case, the SR protein family is typically recognized as exonic splicing enhancers. For example, a particular SR protein could promote exon inclusion by binding to exons and recruiting U1 snRNP to the 5′ splice site and U2 auxiliary factor (U2AF) to the 3′ splice site through protein–protein interactions in early steps of spliceosome assembly [24; 25]. Conversely, the hnRNP protein family was defined by a co-purification experiment that revealed them as abundant nuclear factors [26], and are thus more structurally diverse, containing RRM-, KH, and RGG-type RNA-binding domains as well as low-complexity regions. Lastly, the hnRNP-like family of proteins were identified more recently, and these share sequence similarity to the canonical hnRNPs, but exhibit more tissue-specific and temporal-restricted expression patterns [27; 28].

Splicing Mis-regulation is an Oncogenic Driver. A comprehensive pan-cancer analyses revealed that tumors have on average 20% more alternative splicing events than matched healthy tissues [29; 30]. The regulatory parameters intrinsic to the pre-mRNA splicing process allow for mis-regulation to occur via multiple avenues. Changes to the motif sequence of core splicing signals and SREs are one way by which individual splicing events may be altered. Indeed, an increasing number of genetic mutations found in cancers that have previously been annotated as missense or nonsense actually alter splicing fidelity and affect gene expression by hindering splicing of individual introns and/or generating new splice sites [31; 32; 33]. Conversely, *trans*-acting genetic alterations to RBPs and splicing factors would typically result in wide-spread alterations to RNA-processing patterns. As it happens, many cancers harbor mutations in genes encoding splicing factors, further supporting the fact that disruptions to splicing homeostasis drive cancer progression [22]. Change-of-function mutations have been typically observed in

splice factor genes, thereby changing the sequence specify of the RNA-binding domain and producing subsets of aberrant splicing events. Splice factor mutations are also typically observed to be heterozygous, suggesting that homozygous alterations are not tolerated by the cancer cell.

Alternative mRNA splicing allows for switches in protein isoforms in the absence of permanent changes in the cell's genetic content, and without changes in transcriptional activity [34]. Furthermore, small changes in the level of protein abundance of distinct RBPs are capable of effecting large changes in alternative splicing patterns. Thus, when perturbed, gene expression levels of RPBs and splicing factors are potentially capable of promoting an oncogenic molecular profile independent of genetic alterations. Indeed, SR and hnRNP proteins have been shown to act as oncoproteins when overexpressed in the correct cellular context. For example, modest overexpression of the SR splicing factor 1 (SRSF1; also known as ASF and SF2) is pro-tumorigenic in cancers, including lung, colon and breast cancer [35; 36; 37]. Additionally, downregulation of SRSF1 gene expression has been observed in a subset of AML patients with altered splicing patterns [38], although more work needs to be done to demonstrate oncogenicity. Conversely, other splicing factors may also act as tumors suppressors. For example, *HNRNPK* (among other genes) lies in chromosome 9 locus which is recurrently deleted (-9q) in acute myeloid leukemias, and therefore a portion of patients have significantly reduced hnRNP K protein expression [39]. Mouse studies found that deletion of one *Hnrnpk* allele results in genomic instability and the formation of transplantable hematopoietic neoplasms. The reduced hnRNP K expression promoted proliferation via attenuated p21 activation and activated STAT3 signaling, as well as halted differentiation through downregulated C/EBP levels. [40].

Cancer cells may also leverage alternative splicing to facilitate disease relapse. One study of acute lymphoblastic leukemia (ALL) patients found that alternative splicing of *CD19*, particularly the skipping of exon 2, removed from the epitope targeted by the chimeric antigen receptor immunotherapy (CART-19). The alternatively spliced transcript translated a truncated CD19 peptide that promoted disease relapse [41]. Additionally, the plasticity of RNA splicing allows for the cancer cells to modulate normal splicing programs and generate different transcript variants which mediate chemotherapy resistance. For example, there is evidence that relapsed ALL patient blasts may be skipping exon 11 in *FPGS* to disrupt the function of folylpolyglutamate synthase, an enzyme involved in the intracellular retention of folate antagonist drugs used to commonly used to treat ALL patients [42]. Thus, profiling alternative RNA splicing status within a particular tumor sample unequivocally provides information that aids in the assessment of oncogenicity effected by molecular changes to the transcriptome.

High-Throughput Quantification of Alternative pre-mRNA Splicing. Splicing regulation has been traditionally studied either at the level of full gene isoforms or through the specification of alternative splicing 'events'. The elucidation of splicing regulation has benefitted tremendously from experimental methods such as in vitro splicing assays [43]. However, the sheer complexity of splicing regulatory mechanisms warranted the development of high-throughput methods that accurately map transcriptome-wide variations in alternative pre-mRNA splicing. In particular, RNA-sequencing (RNA-Seq) technology produces millions of read sequences derived from gene transcripts that can be leveraged to produce models of alternative splicing events. Splicing quantification

algorithms utilize read sequences that span over splice junctions to create a 'splicegraph' that models the inclusion ratios of a particular splicing event.

Determining whether two distant exons originate from the same mRNA transcript is a problem that can be addressed by long-read sequencing technologies (e.g., PacBio, Oxford Nanopore). However, long read sequencing is currently very costly, and is unable to properly resolve small variations (<10nt) between isoforms, highlighting detection limits of the technology [44]. Also, sequencing errors, alignment and amplification artifacts posing as novel splice sites in long-reads present a challenge for many existing splicing quantification tools, and thus studies have supplemented noisy long-read sequencing data with short-read sequencing data to rescue reads with incorrect splice sites and facilitate downstream analysis [45]. As discussed in the previous section, the maturation of pre-mRNAs requires the polyadenylation of the 3' end of the transcripts. Thus, sequencing of polyadenylated RNA provides a robust catalog of transcripts expressed by most genes with known function. The downside to using short-read RNA-seq data include potentially limited coverage depth across genes expressed at low levels as well as bias towards particular regions of the transcript. Specifically, oligo-dT priming and DNase fragmentation are known to generate read bias towards the 3' end of transcripts. However, despite these challenges, short-read RNA-seq data produces a robust catalog of molecular abundance for a particular gene transcript library which can be parsed to produce quantitative models of alternative splicing (**Figure 1.3**).

Exon–exon junctions can be identified by the presence of the GT–AG dinucleotides that flank splice sites and confirmed by the low expression of intronic sequences, which are removed during splicing. A variety of computational methods for the high-throughput quantification of alternative pre-mRNA splicing have been developed for short-read RNA-

seq analysis, such as SUPPA [46], rMATS [47], MISO [48], SpliceTrap [49], rSeqDiff [50], Cufflinks [51], FDM [52], DiffSplice [53], DEXSeq [54]. Initially, these algorithms had critical limitations in detecting differential splicing and/or isoform proportion while accounting for variability among replicates and could not handle large sample sizes. Initial algorithms were also only capable of quantifying annotated binary splicing events. While useful, it is evident that binary alternative splicing events fail to capture the full complexity of spliceosome decisions [17]. To address the shortcomings of initial high-throughput splicing modeling paradigms, Vaquero et al. developed the MAJIQ model (Modeling Alternative Junction Inclusion Quantification), a framework that efficiently combines RNA-Seq data with existing gene annotation and to enable the accurate detection, quantification, and visualization of complex splicing variations across large numbers of experimental conditions.

Unlike many of the aforementioned methods that only analyze known isoforms, MAJIQ supplements known transcripts with *de novo* junctions inferred from the RNA-seq data. The results of MAJIQ's splicing quantification can be interactively visualized with the VOILA package in a standard web browser. Importantly, the MAJIQ model quantifies splicing events in terms of *local splicing variations*, or LSVs, that individually represent substructures in the transcriptome. LSVs capture *complex* alternative splicing events (**Figure 1.2C**), which were found to be more prevalent in human transcriptomes (over 30%). An LSV is defined as a single exon alternatively joined to more than one RNA segment. LSVs are visualized as splits (multiple edges) in a splicegraph where several edges either come into or from a single exon, termed the reference exon. Junctions involved in an LSV are quantified in terms of Percent Splicing Index (PSI, Ψ), a unit value that represents the relative fraction of poly-A selected mRNA transcripts in which the

aforementioned exon is joined to each of the alternative RNA segments (**Figure 1.3**). MAJIQ can quantify LSVs within a single sample, as well compare two experimental conditions with or without replicate groups through differential splicing analysis across large heterogenous datasets. LSV quantification in a single condition is based on the marginal PSI (Ψ) for each junction involved in the LSV, while the differential comparison of experimental conditions is based on relative changes in Ψ (dPSI, ΔΨ). The MAJIQ framework uses Bayesian modeling to report posterior Ψ and ΔΨ distributions for each quantified LSV (**Figure 1.3D**). Importantly, MAJIQ empirical PSI estimates inferred from RNA-seq data have achieved robust overall correlations (R_{Ψ} = 0.8, $R_{\Delta\Psi}$ = 0.95) with splicing quantification done by radiolabeled RT-PCR in an analysis of datasets from multiple tissues. Moreover, MAJIQ's reproducibility of both binary and complex LSVs is stable across varying read coverage depth, further underscoring the robustness of the method. Notably, MAJIQ analyses have also indicated that including *de-novo* junctions increases the number of differentially spliced LSVs that could be detected by around 30%. Therefore, the reproducibility, accuracy and LSV detection power of the MAJIQ algorithm, as well as its ad-hoc visualization tools, make a computational framework suitable for the analyses of mis-regulated complex transcriptomes intrinsic to cancer.

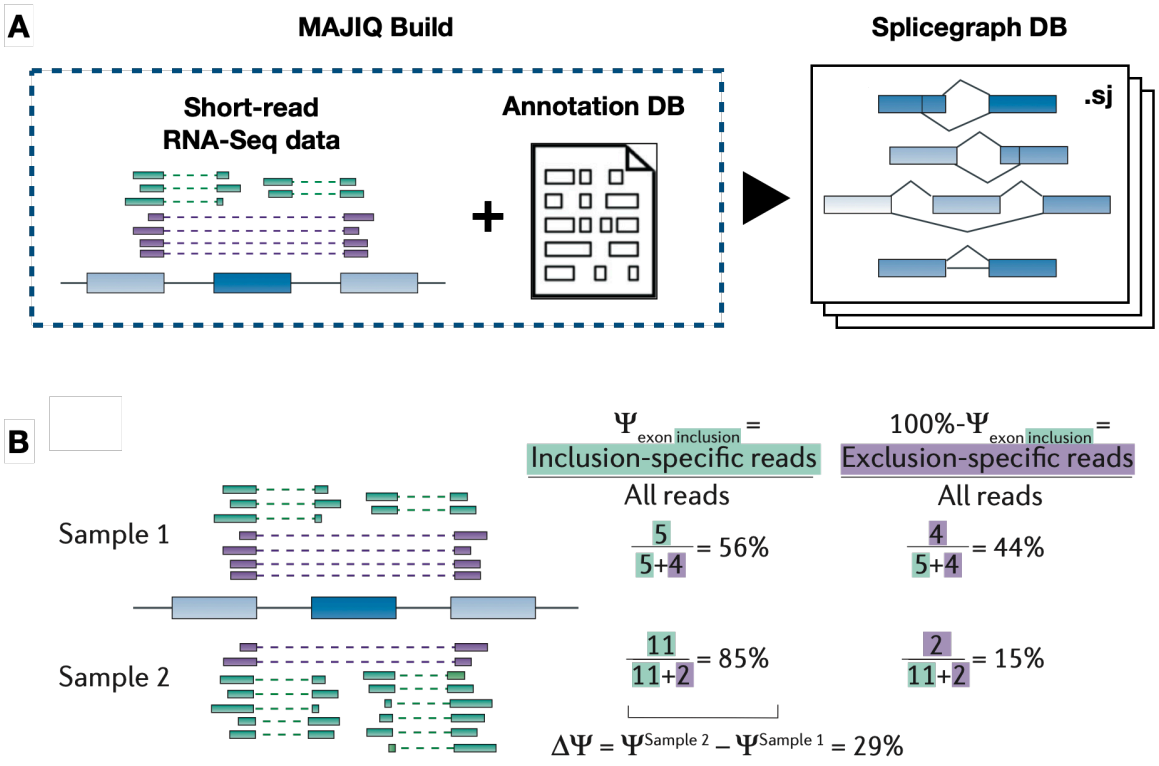


Figure 1.3 – **High-Throughput Quantification of Alternative Splicing.** (**A**) Schematic of MAJIQ Splicing Event Identification and Quantification. RNA-seq data in the form of aligned and sorted bam files is submitted alongside a genomic database of annotated splicing events. The MAJIQ Build algorithm produces a collection of .sj files that contain all of the quantified junctions for a particular sample alongside a splicegraph database that collectively allows for the visualization of the splicing models that included splicing decisions across all of the selected samples. (**B**) Quantification of Percent Splice Index (PSI) using junction-spanning reads. The ratio of inclusion specific reads for a particular junction is denoted by (PSI, Ψ). Differential inclusion of junctions is calculated between samples and is denoted as deltaPSI (ΔΨ) (Adapted from Dvinge et al. 2016)

Acute Myeloid Leukemia

AML disease heterogeneity. The formation of blood cellular compartments – namely hematopoiesis – is a process settled in the bone marrow driven by a rare subpopulation of self-renewing hematopoietic stem cells (HSCs). The molecular mechanisms driving the proliferation and terminal differentiation of hematopoietic cell populations are mostly conserved across mammals [55]. HSCs are endowed with regulated quiescence and self-renewal capabilities that preserve the multipotency of dividing cells and sustain the generation of adequate levels of blood cellular compartments over a human's lifetime [55; 56; 57]. When this delicate balance of differentiation, proliferation, and self-renewal is disrupted, it results in the development of hematopoietic disorders. Leukemias are liquid tumors that specifically arise from the transformation of hematopoietic stem cells and progenitor cells in the bone marrow. Thus, the malignant transformation of healthy bone marrow cells is a multistep process that involves the acquisition of molecular lesions that collectively impede normal cell programs and result in the generation of neoplastic clones with dysregulated developmental and proliferative properties [58; 59].

Acute myeloid leukemia (AML) is a class of hematologic cancer characterized by the aggressive clonal proliferation of abnormally differentiated myeloid precursor cells (**Figure 1.4**) [60]. According to World Health Organization (WHO) Classification system guidelines, the diagnosis of AML requires that myeloblasts constitute more than 20% of the bone marrow or circulating peripheral blood white cell population [61]. Additionally, the percentage of blood cell compartments such as promyelocytes, monoblasts, promonocytes and megakaryoblasts are included in the classification process, and are used to define the various AML subtypes. Cells obtained from AML patients – namely AML *blasts* – often have striking morphological similarities to normal HSCs and often

share similar cell surface marker profiles. Monoclonal antibodies targeting cluster of differentiation (CD) cell surface proteins are restricted to cells committed to myeloid differentiation (e.g., CD11b, CD13, CD14, CD33), but are found in untransformed cells and are thus not representative of a leukemia-specific prognostic signature or unique to different AML subtypes, with the possible exception of CD34 [62]. Specifically, CD34 is detected on undifferentiated hematopoietic progenitors and can also be found on the blasts of patients with either AML or ALL. There is evidence that patients with AML whose blasts strongly express CD34 have an inferior outcome due to chemotherapy-resistant leukemia [63]

Historically, AML blasts have been studied through the lens of distinct chromosomal rearrangements and translocations that appear to drive the oncogenicity of particular subtypes of leukemia [64]. The different translocations initially observed in leukemias were generally found stably within the entirety of the malignant clonal patient cell population, thus it was thought that these molecular lesions occur as early *hits* during disease development and have a direct causal role in disease progression and pathogenesis [58; 65]. Thus, several of these recurrent chromosome abnormalities are now included in the 2008 revision of the World Health Organization (WHO) classification scheme of AML. For example, translocation t(15;17)(q22;q12) generates the *PML–RARa* fusion gene and is exclusively associated with acute promyelocytic leukemia (APL, AML–M3) [66]. Cytogenetic analysis of AML patient blasts has demonstrated clinical efficacy in subsets of patients with clonal lesions that have proven chemotherapeutic agents [67; 68]. However, it is now understood that genetic mutations such as single nucleotide substitutions and small tandem duplications account for more than 70% of all driver molecular lesions in AML [60].

Genetic mutations typically co-occur with chromosomal rearrangements, but also arise independently of any karyotypic abnormality. Importantly, almost half of all AML patients have a cytogenetically normal karyotype (CN-AML) and has been historically classified as intermediate risk according to the European Leukemia Network (ELN) guidelines. Interestingly, although patients typically harbored more than 1 driver genetic mutation, particular sets of these genetic features have been able to cluster AML patients into groups with significant prognostic relevance [69]. Furthermore, studies suggest that distinct genetic abnormalities have a likelihood of occurring between early and late disease development [70; 71]. Therefore, AML blast clones have the potential of evolving over time, further acquiring genetic lesions and generating multiple competing clones at any time within the same patient (**Figure 1.4**). Consequently, this phenomenon results in negative outcomes such as lower overall survival, chemotherapeutic resistance and disease relapse. Thus, while most AML patients without cytogenic abnormalities have been historically classified as intermediate risk, most patients indeed harbor genetic mutations that could confer aggressive oncogenic cellular programs. Therefore, the current gold-standard of care for AML patients involves performing testing for recurrent mutations in genes associated with AML, in conjunction with morphological and cytogenic analysis, to comprehensively diagnose the patient disease state and create a personalized treatment model.

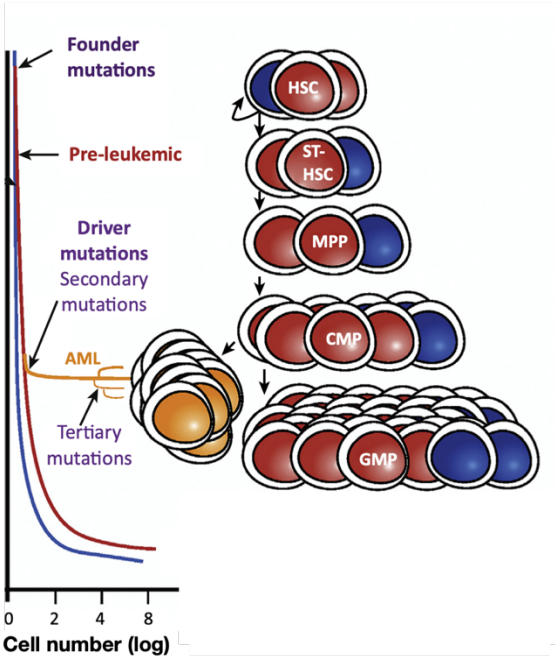


Figure 1.4 – AML Heterogeneity. Schematic representing how founder mutations in AML-associated genes lead to the establishment of a preleukemic hemopoietic stem cell population that upon incidence of secondary driver mutations develop into leukemia. The additional accumulation of multiple tertiary mutations results in the patient-patient molecular heterogeneity. Intrinsic to the AML population. Hematologic progenitor cells are denoted as such HSC/MPP/CMP/GMAP (Adapted from Brumatti et al. 2017)

AML-associated genes. Comprehensive mutational analysis of AML patient cohorts has revealed sets of genes that have recurrent genetic mutations (**Figure 1.5**). Interestingly, mutated genes span a range of categories such as hyperactivated growth signaling, hematopoietic transcription factors, chromatin modifiers, DNA methyltransferases and RNA splicing factors [60]. Particular genetic mutations identified within AML-associated genes express higher penetrance than other commonly studied cytogenic abnormalities and chromosomal rearrangements. Furthermore, these genetic mutations exhibit complex patterns of cooperation and mutual exclusivity between themselves, suggestive of the need for distinct additive effects that are balanced to promote the oncogenicity of AML blasts.

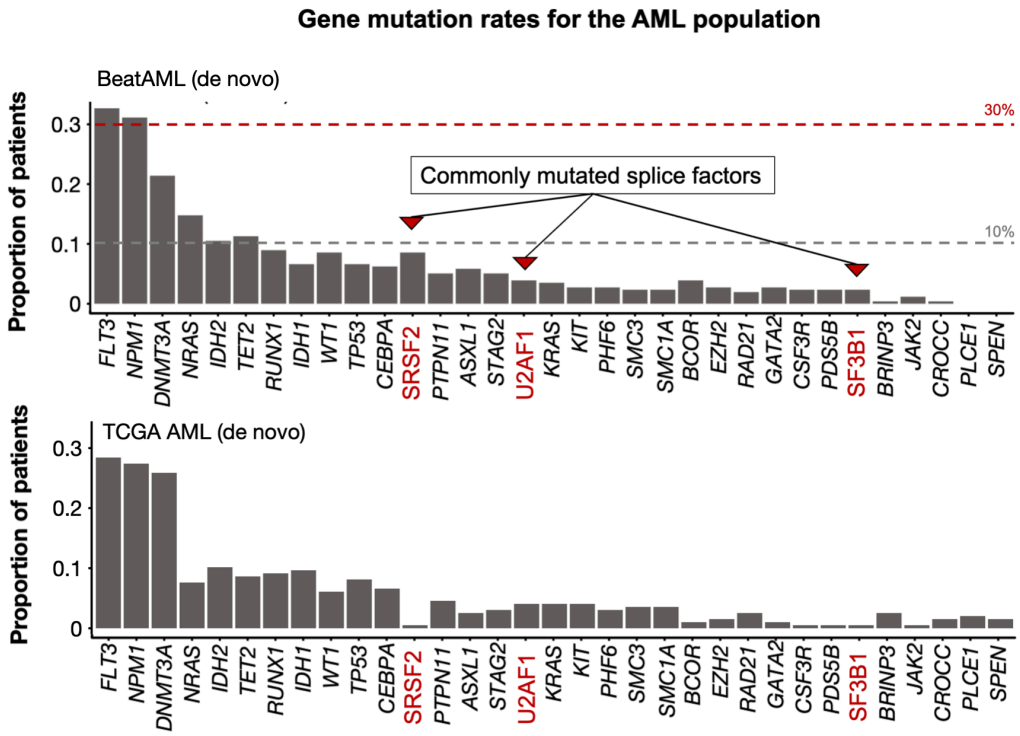


Figure 1.5 – Frequency of mutational events in AML-Associated genes that are cumulatively the most frequent across two independent patient cohorts of *de novo* AML – BeatAML (n = 531) and TCGA-LAML (n = 200 patients) (Data obtained from Tyner et al 2018 Figure 1A) Of note, although only one patient with a mutation in *SRSF2* was reported in the TCGA-AML publication, retrospectively mutational analysis of RNA- seq data identified SRSF2 hotspot mutations in each of these 19 patients (11% of the patients with AML) [72]. This retrospective finding is consistent with the frequency of mutations in SRSF2 identified across patients of the BeatAML cohort.

The gene *FLT3* (Fms Related Receptor Tyrosine kinase 3, also known as CD135) is the most commonly mutated gene across the AML population, with around 33% of patients harboring an internal tandem duplication (ITD) that is not identifiable through traditional cytogenetic analysis [73]. The FLT3-ITDs are nucleotide duplications whose length can vary from 3 to more than 400 base pairs generally recurring between exons 14 and 15, resulting in the insertion of repeated amino acid sequences within the juxta-

membrane domain of the FLT3 protein. The encoded FLT3 transmembrane protein is selectively expressed in CD34+ hematopoietic stem cells [74] and its coordinated signaling is important for proper development of normal hematopoietic stem cell self-renewal and differentiation. Crystallography studies showed that FLT3 juxta-membrane domain makes contact with the activation loop to maintain the protein in an auto-inhibited monomeric state [75]. The kinase activity of FLT3 is normally stimulated upon ligand binding, which promotes the dimerization and the concomitant juxtaposition of the cytoplasmic domain of the FLT3 receptor and facilitates trans-phosphorylation of its tyrosine residues [73; 76]. Thus, the ITDs in the FLT3 juxta-membrane domain as well as mutations in the activation loop promotes hypersensitive receptor dimerization as well as ligand-independent trans-autophosphorylation. Consequently, the constitutive activation of the FLT3 signaling pathway results in cytokine-independent proliferation and blocking myeloid differentiation of hematopoietic progenitors [77]. Consistently, increased allelic ratio of the FLT3-ITD mutation is known to correlate with poor prognosis. [78]

The second largest AML disease subgroup, accounting for slightly less than 30% of the population, is defined by mutations in the gene *NPM1* (Nucleophosmin 1). The gene encodes a molecular chaperone phosphoprotein that oligomerizes to form a pentamer, and sometimes a decamer, that shuttles between nuclei and cytoplasm, with pleiotropic functions including ribosomal biogenesis [79], centrosome duplication [80], and regulation of p53 [81]. Mutations, typically insertions, in the *NPM1* gene are typically concentrated in exon 12 and involve a frameshift (W288fs) which results in the cytoplasmic sequestration of a truncated protein. Interestingly, mice expressing the *Npm1*-W288fs mutation develop myeloproliferative neoplasms but not overt leukemia [82], suggesting that it may require additional mutations to promote leukemic development. Concordantly, many patients

carrying *NPM1* mutations also carry the FLT3-ITD mutation as well as mutations in one or more DNA methylation genes such as DNMT3A [64]. *NPM1* mutations in the absence of FLT3-ITD or other mutations are associated with good prognostic outcomes.

Of note, mutations in NPM1 were significantly exclusive of point mutations that occur in hematopoietic transcription factor RUNX1 [69; 83]. The RUNX family transcription factor 1 (RUNX1) is essential for the function of hematopoietic stem cells and is the most frequent target for chromosomal translocations in leukemia [83; 84]. Loss of function point mutations are mostly clustered in the Runt domain at arginine residues R107, R162, and R201 and aspartate residue D198. In particular, biallelic mutations in *RUNX1* have been observed in AML-M0 patients, indicating a complete lack of RUNX1 function in their leukemic cells [85]. Therefore, leukemogenesis in RUNX1 mutated patients is thought to promote growth advantage to the hematopoietic progenitor cells with differentiation defects that result from mutations in other genes [83]. Intriguingly, mutations in RUNX1 significantly co-occurred with mutations in DNMT3A, chromatin factors and splicing factors [69]

Around 20% of AML patients harbor point mutations in DNA methyltransferase 3 Alpha (DNMT3A), which has been recently described as a recurrent founder mutation [69]. Further analyses found mutations in genes that encode other epigenetic modifiers such as *ASXL1* and *TET2*. In particular, DNMT3A catalyzes the addition of a methyl groups to concentrated clusters of cytosine residues in DNA regions upstream of genes, thereby reducing the expression of the downstream gene [86]. Aberrant DNA methylation is widely accepted to contribute to the pathogenesis of AML and of cancer overall. Of note, clonal patterns with initial mutations in these epigenetic factors may potentially promote the

occurrence of myelodysplastic syndrome (MDS), a pre-malignant disease that also affects myeloid cells and is known to progress to AML

Additional epigenetic avenues such as chromatin remodeling are disrupted in AML. Partial tandem duplications (PTDs) in Lysine Methyltransferase 2A (KMT2A also known as Mixed Lineage Leukemia 1 or *MLL1*). The KMT2A protein SET domain is responsible for histone H3 lysine 4 (H3K4) methylation and is implicated in regulating the transcription of specific target genes, including the *Hox* gene cluster which are involved in hematopoiesis [87]. Loss of function point mutations in other chromatin factors such as EZH2 particularly downregulate the catalytic activity of the SET domain which is responsible of deposition of di- and trimethylation on lysine 27 of histone H3 (H3K27me2/3) [88]. Loss of EZH2 protein induces resistance to multiple drugs in AML, presumably through derepressing the *HOX* gene cluster [89]. Overall, mutations within hematopoietic transcription factors, signal transduction receptors, epigenetic regulators and chromatin modifiers broadly dysregulate networks of gene expression, and larger-scale studies of AML population have sought to determine how gene expression profiles confer oncogenic capacity of the malignant blasts

Around 18%-20% of AML patients harbor mutations in genes encoding RNA splicing factors U2AF1, SRSF2, SF3B1, and ZRSR2, which are all involved in 3' splice site selection (**Figure 1.6**) [21; 90]. Importantly, mutations in these splicing factors are mutually exclusive, suggesting that two mutations in distinct splice factors is not viable for the cell. Mutations in splice factors in particular showed slight patterns of co-occurrence with mutations in chromatin factors as well as RUNX1 and other myeloid transcription factors, but were significantly exclusive of mutations in DNMT3A and NPM1 [91]. The SR-protein SRSF2 – previously known as SC35 –binds to exonic splicing enhancers and is

required for the ATP-dependent interactions between U1 and U2 snRNP. Specifically, hotspot mutations in SRSF2 typically convert residue proline 95 (P95) to a histidine (H), leucine (L), or arginine (R) which alters the RNA-binding specificity of the RNA-binding domain and drives preferential recognition of cassette exons containing C-rich versus G-rich sequences in exon exonic splicing enhancers (**Figure 1.6A**) [92]. Inducible Mx1-Cre Srsf2P95H/WT knock-in mice displayed hallmarks of myeloid disorders such as MDS [92].

Similarly, the SR-protein U2AF1 also harbors hotspot mutations, which fall in the Zn knuckles and typically affect either residues S34/S35 or R156/Q157 and are also thought to alter preferential binding to the 3' ss (**Figure 1.6B**). U2 auxiliary factor 1 protein – also known as U2AF35 – is 65 kaDa subunit non-snRNP protein required for the binding of U2 snRNP to the pre-mRNA branch site. Recombinant mice conditionally expressing U2AF1(S34F) had impaired hematopoiesis but did not develop leukemia, suggestive of the need for cooperative mutations, particularly in hematopoietic transcription factors [93]. Additionally, mutations in spliceosome machinery, usually within U2AF1, have been found to increase the occurrence of MDS with high probability of development into AML [94]

Mutations in splicing factors SF3B1 and ZRSR2 are significantly less frequent in AML (**Figure 1.6C-D**) [69,95]. The SF3B1 protein – previously known as SAP155 –is the largest component of a 450-kDa hetero-heptameric SF3B complex, a subunit of 17S U2 snRNP and the analogous minor spliceosomal U12 snRNP [96]. The splicing factor 3b/3a complex binds pre-mRNA upstream and downstream of the intron's branch point sequence and recruits the U2 snRNP to the pre-mRNA [97]. In particular, point mutations in SF3B1 mostly occur at lysine residue 700 (K700), which falls within the HEAT domain of SF3B1 [90]. Mutations in this residue disrupt the interaction with another auxiliary

splicing factor, SUGP1, and interfere with the interactions that support the scaffolding for the U2 snRNP and other SF3B subunits, resulting in widespread changes in alternative splicing [98]. Similarly, mutations in ZRSR2, which are loss-of-function missense, frameshifts and nonsense mutations that occur across the protein and are also thought to interfere with protein-protein interactions with other components of the spliceosome [99]. ZRSR2 is known to interact with the U2AF1/2 complex and also regulate the splicing of introns U12-dependent minor spliceosome [100; 101].

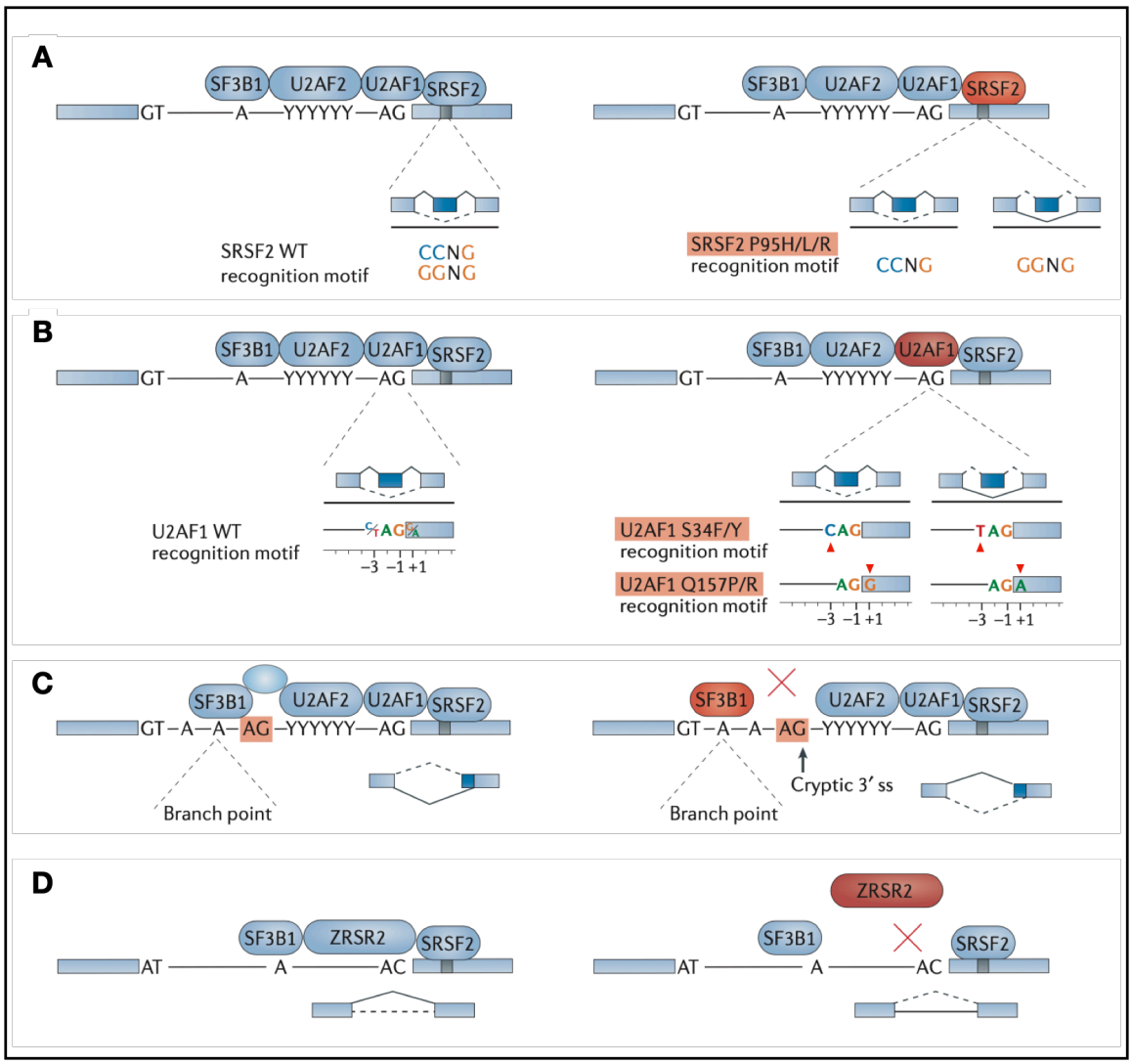


Figure 1.6 – RNA Splicing Factor Mutations in AML (**A**) SRSF2 hotpot mutations P95H/L/R alters the RNA-binding domain causing preference for an exonic enhancer with CCNG motif over a GGNG motif. (**B**) U2AF1 hotspot mutations at residue S34 and Q157 create distinct preference for cysteines and guanines at the 3' splice site respectively (**C**) Mutations in SF3B1 interfere with the spliceosome protein scaffolding and impede branch point recognition which may lead to cryptic splice site usage (**D**) Mutations in ZRSR2 also interfere with protein-protein interactions between spliceosome machinery and may induce aberrant retention of U12-type introns. (Adapted from Dvinge et al. 2016 Figure 3)

Transcriptomic Profiling of AML Patients. A handful of researchers have previously sought to define dysregulated gene expression and splice site choices associated with the leukemogenesis of AML blasts. Historically, exploratory studies of the AML transcriptome were limited to Affymetrix microarray gene chips [102; 103; 104; 105] and exon junction probe sets [106; 107] to quantify changes in gene expression and alternative splicing respectively. For example, microarray studies have identified overexpression of the *HOX* gene cluster in particular within AML patients with *NPM1* mutations as well as with +8/8q karyotype. More recently, studies utilizing high-throughput short-read RNA sequencing have sought to further identify prognostic signatures in gene expression and alternative splicing variations. For example, unsupervised clustering analysis of RNA-seq expression data from at least one large AML cohort revealed correlations with overall survival, as well as with particular morphologically defined subgroups of AML and genetic mutations in AML-associated genes [64]

The overexpression of genes pertaining to the *HOX* gene cluster has also been found implicated in leukemogenesis, AML disease progression and overall poor prognosis [108; 109; 110]. Interestingly, one study in pediatric AML found it possible to distinguish patients with NPM1 mutations from those with 11q23/MLL translocations based on different patterns of *HOX* gene expression [111]. Additionally, one study performed in recombinant Mx1-cre mice linked the expression of several hematopoietic regulators including upregulation *Hoxb2* and downregulation genes such as *Gata1* and *Gata2* to the conditional expression of *Srsf2P95H* [92]. Alternative pre-mRNA splicing promotes oncogenicity via *de novo* synthesis of aberrant splice isoforms as well as through disruption of normal ratios of splice isoforms. For example, it was demonstrated in mouse models that alternative splicing of *Hoxa9* translates to a truncated version of the protein

(Hoxa9T) which binds strongly to CREB-binding protein and activates transcription of proto-oncogenes [112].

Transcriptomic perturbations to only some of the AML-associated genes have been previously described (See Table 2.1). FLT3 is one of the most highly mis-spliced genes in AML with aberrant splicing of the transcript primarily altering an extracellular region of this receptor [113]. The FLT3 pathway frequently hyperactivated in AML controls many downstream target genes, including apoptotic regulator *BIRC5*. Splice variant ratios of the *BIRC5* gene, which encodes the Survivin inhibitor of apoptosis, are significantly associated with risk of refractory disease and poor survival outcomes in both children and adults [114; 115]. The gene *RUNX1* also frequently mutated in AML encodes a minor splice isoform often seen at higher levels in AML patients that lacks the transactivation domain and binds DNA with higher affinity than the full-length isoform thereby acting in a dominant negative fashion [116]. Furthermore, the handful of studies that have sought to uncover differential splicing changes that are common across AML patients usually compare an entire group of patients to non-malignant cells [117;118;106]. However, this particular strategy overlooks the intrinsic heterogeneity of AML given that the assumption with this model is that every AML patient has the same molecular background, and this has been demonstrated to be untrue. The aforementioned splicing studies also use algorithms that fail to capture *de novo* and complex splicing variations, thus overlooking important changes that likely arise in the cancer transcriptome.

It is evident that analysis of mutational status level alone underestimates the molecular heterogeneity of AML blasts and overlooks important variations at the level of pre-mRNA processing and gene expression. Importantly, variations in pre-mRNA splicing within AML-associated genes themselves are capable of eliciting loss-of function or change

of function effects. However, transcriptome-wide splicing quantification analyses have also only been done centered around a genetic mutation in a splicing factor or a particular AML-associated gene. Thus, an alternative question, which has not been addressed by previous studies, is whether splicing variations across AML patients serve as an independent mechanism to somatic mutations by which the function of AML-associated genes may be altered. In the work performed for this dissertation, I investigate how pre-mRNA splicing variations functionally dysregulate AML-associated genes. The overarching goals of this dissertation are to expand the repertoire of transcriptomic variations expressed among the AML population and to explain the mechanisms by which an mRNA splice variant can phenocopy the effects of genetic mutations within AML associated genes. In the following chapters, I elaborate on an exploratory analysis deployed over an in-house cohort of AML patients. The analyses lead to the identification of alternative splicing patterns that were also observable in a larger and independent AML cohort with a striking degree of correlation. The occurrence of a handful the most interesting splicing events was validated experimentally in patient derived AML blasts. Finally, I expand the landscape of commonly altered genes in AML by identifying deleterious splicing events within genes not previously linked to the development of this disease. Taken together the body of work presented in this dissertation underscores that future studies and clinical pipelines need to account for variations in mRNA splicing to fully profile the molecular heterogeneity of AML and design personalized therapeutic strategies.

CHAPTER II

Alternative splicing functionally disrupts genes associated with AML

Introduction

No previous studies have robustly profiled variations in RNA splicing within genes that occur with high mutational frequency in AML. As discussed in Chapter 1, splicing variations may serve as an additional mechanism to somatic mutations by which the function of these genes may be altered. Specifically, for the studies discussed in the following chapter, I hypothesized that accounting for variations in splicing would increase the percentage of AML patients that harbor molecular alterations within a particular AML-associated gene. To test the above hypothesis, I quantified splicing variability of genes that have been identified as commonly mutated, truncated, or translocated in AML patients. The geneset was initially obtained from the analysis of genetic mutations performed as part of the TCGA-LAML project [64], which was among the first large scale studies to perform coupled whole genome and exome sequencing analysis in AML to identify candidate somatic variants that a recurrently associated with AML. An additional study of AML genomes, the German-Austrian AML study group, which spanned multiple centers, confirmed the genetic mutations described by the TCGA study for an *a priori* defined set of genes [69]. Furthermore, a yet another large study of AML genomes confirmed most of the previously identified mutations [95] and a retrospective analysis of RNA-seq data of TCGA-AML highlighted the overlooked increased occurrence of SRSF2 mutations in AML [72].

For my analysis of splicing variations in AML-associated genes, I compiled a list of those genes that were found to be commonly mutated in more than 1.5% of AML patients across the aforementioned studies. The curation resulted in a list of 70 genes of interest that I focused on for the analyses discussed in the following Chapter. Herein, I analyze polyA-selected RNA-Seq data from 29 in-house de-identified AML patient samples

collected by the Penn Stem Cell and Xenograft Core (aka PENN-AML cohort). Samples in the PENN-AML cohort were isolated by leukapheresis or from peripheral blood mononuclear cells (PBMCs), and RNA-Seq data was generated from samples with an AML blast purity of greater than 90% (sample information published in Rivera et al 2021, RNA-Seq data is available in GSE142514). Thus, samples within the PENN cohort can be quantitatively compared without concern of confounding factors tied to sample purity. Specifically, I used the novel MAJIQ computational framework discussed in Section 1.4 to process the RNA-sequencing files from the PENN-AML patient cohort (**Figure 2.1A**). By using the MAJIQ algorithm, the analysis was optimized to detect and quantify complex as well as novel splicing events, and thus was not limited to previously observed or expected splicing patterns.

As previously discussed in Chapter 1, the MAJIQ framework outputs a catalogue of local splicing variations (LSVs) that represent any splicing events, constitutive or variable, that are quantified transcriptome-wide or for a defined set of genes. Given the context of my study, I particularly focused on configuring the MAJIQ framework to quantify splicing events for the aforementioned 70 AML-associated genes (**Figure 2.1B, C**). The LSVs that are outputted by MAJIQ represent a highly dimensional dataset with some level of redundancy. In other words, there are cases where LSVs overlap with one another and point to the same splicing event but through different angles of the transcript. To resolve these cases and to reduce the dimensionality of the LSV dataset, I coded a pipeline that groups overlapping LSVs into *splicing modules* (**Figure 2.1D**). Thus, each splicing module represents either 1 individual LSV, which in turn represents 1 splicing event, or 2 overlapping LSVs that share a particular junction of what is actually the same single splicing event. If two LSVs formed a splicing module, the most variable junction across all

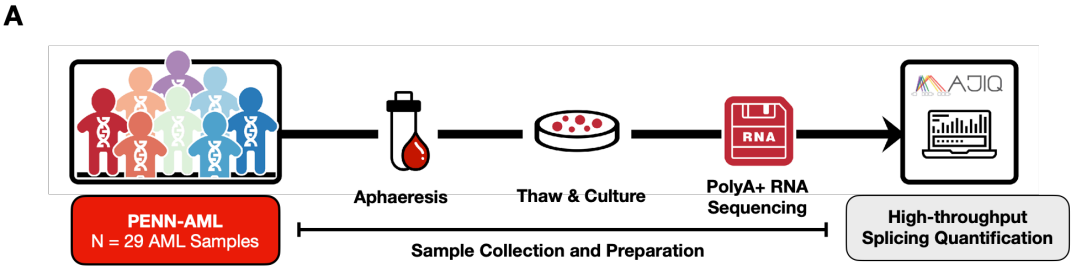
the grouped junctions in both LSVs was selected as the junction representative of the variance for the particular module.

The goal of this study is to quantify splicing variation that underlies patient-to-patient heterogeneity; therefore, of interest were the most highly variable splicing modules quantified within the dataset (**Figure 2.1D**). Therefore, for each module, the pipeline calculated the variance in PSI for each junction, which is essentially how much the percent usage of each junction deviates across the sample cohort. For each splicing module, the most variable junction within the splicing module was selected as the representative junction. The splicing module data feature facilitates downstream analyses by focusing on using only the values that are representative of the underlying biological variation in question across AML cohorts. Upon identifying a particular splicing module of interest, I can query the outputted database for the particular underlying LSV and visualize it in the VOILA visualization engine that is part of the MAJIQ framework. Herein, I provide a catalogue of the variations in pre-mRNA splicing many of which are novel and have not been described in previous studies (**Table 2.1**). Furthermore, I find that AML patients harbor highly variable splicing events with a striking level of correlation. This finding fits well with the widely accepted fact that alternative pre-mRNA splicing events are known to be co-regulated by distinct RBPs.

In my analysis of splicing variations, I also included the BeatAML cohort, a large multicenter study that performed RNA-seq analysis in 444 AML patients [95]. Importantly, as I will discuss in detail in the following Chapter, the splicing events that I uncover in the PENN-AML cohort are also found to behave similarly within the BeatAML cohort. Additionally, I experimentally validate a handful of splicing events of interest using RNA isolated from a subset of patients that form part of the PENN-AML cohort. The correlation

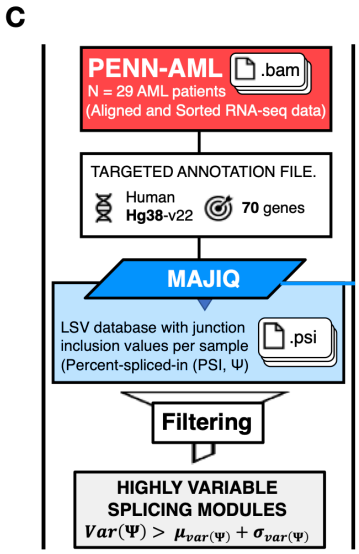
between the experimental results and the MAJIQ-PSI quantifications were particularly striking and thus serve to highlight the power of the MAJIQ computational framework for the robust detection of molecular alterations at the level of mRNA splicing within patient RNA-seq data. Therefore, the results discussed in the following Chapter strongly support that analysis of mutational status alone underestimates the true percentage of AML patients exhibiting functional downregulation of a particular gene. Therefore, my study for the first-time sheds light on dysregulated splicing variations that may be indicative of particular subsets of AML patients. Through the body of work discussed in this Chapter, I aim to motivate studies that address the contribution of pre-mRNA splicing variations to the pathology of AML.

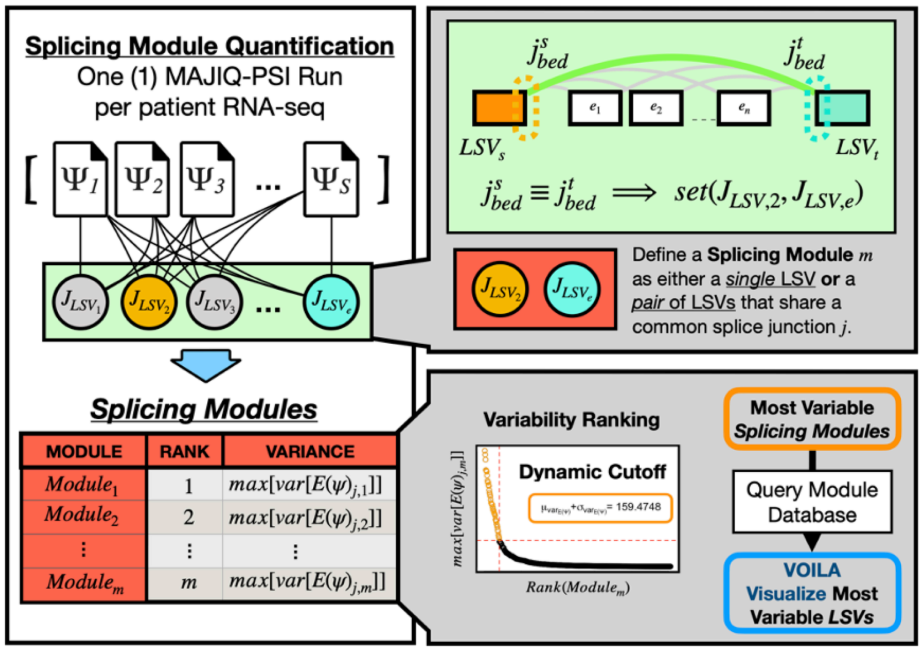
Figure 2.1 – Study design and splicing quantification pipeline (**A**) Workflow of transcriptomic analysis from 29 primary AML samples obtained from the Penn Tumor Bank. (**B**) List of 70 AML-associated genes targeted in MAJIQ analysis. (**C**) MAJIQ pipeline for the analysis of splicing variations of AML-associated genes within PENN-AML patient RNA-seq data (**D**) Schematic of splicing module definition using extended pipeline coded for the purposes of this study.



B

AML-associated genes				
ABL1	FLT3	NPM1	TET2	MLLT3
ASXL1	GATA2	NRAS	TP53	MYH11
ATM	GNAS	PHF6	U2AF1	NSD1
BCOR	HNRNPK	POT1	WT1	NUP98
BCORL1	IDH1	PTEN	ZRSR2	PICALM
BRAF	IDH2	PTPN11	BRINP3	PML
CALR	JAK2	RAD21	SMC3	RARA
CBL	KIT	RUNX1	BCR	RUNX1T1
CEBPA	KRAS	SETBP1	CBFB	SAXO2
CSF3R	MPL	SF3B1	ELL	FLG
DNMT3A	MYC	SMC1A	TMEM255B	KCNU1
ETV6	MYD88	SRSF2	PDGFRA	MYOM3
EZH2	NF1	STAG2	KMT2A	PKHD1L1
FBXW7	NOTCH1	TBL1XR1	MLLT10	PPP1R3A





Results

Highly variable splicing events within AML-related genes. Splicing modules that exhibited a variance greater than one standard deviation above the mean variance for all modules were deemed as the '*highly-variable*' splicing modules within AML-associated genes. (**Figure 2.1A**). The MAJIQ pipeline quantified a total of 506 splicing modules that were quantifiable in at least 80% of the PENN cohort (at least 24/29 patients). Most of these splicing events exhibit low variability amongst samples which is expected given the general need for splicing to be conserved as a consistent step in gene regulation. However, I identified 40 splicing modules that poses a variance above the calculated threshold cutoff (**Figure 2.2A-B**). These 40 highly variable splicing modules implicated 25 of the 70 AML-associated genes. Multiple splicing modules within the same gene suggests that the occurrence of multiple distinct avenues of biological regulation. Importantly, gene expression for these same 70 genes showed overall less variability than splicing, and none of the genes harboring highly variable splicing modules were amongst those with highly variable expression. Importantly, splicing variability of all of the 40 highly variable modules is independent of differences in transcript abundance and *cis*-acting mutations. Furthermore, although most of these splicing modules are predicted to have deleterious effects on the expression of the encoded protein, most of these splicing modules have not been described in the literature (**Table 2.1**). The observed variability in splicing that is independent of gene expression patterns and mutational status across PENN-AML patients underscores significant differences in the AML transcriptome that are not captured by standard genetic and gene expression analysis.

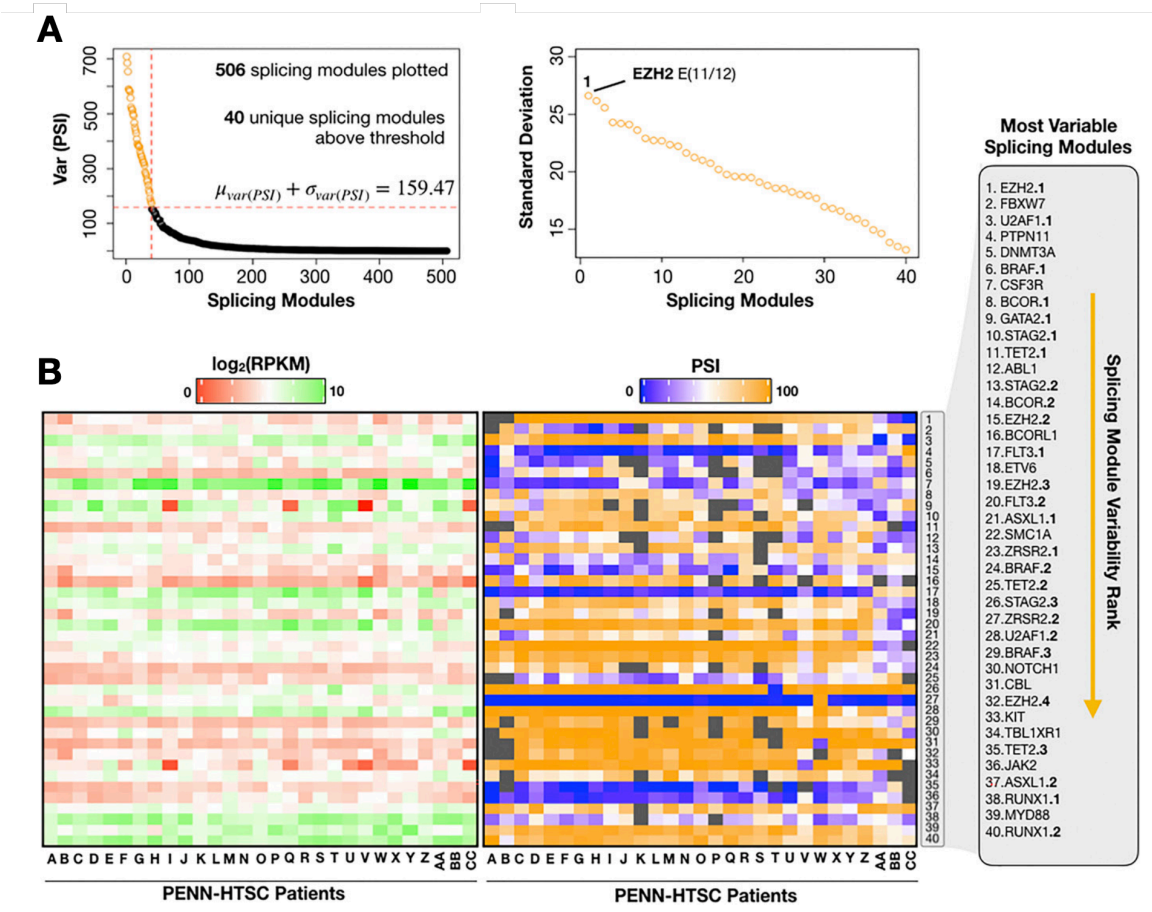


Figure 2.2 –**Splicing variability and gene expression heterogeneity of AML-associated genes across the PENN-AML patient cohort.** (**A**, left) Distribution plots of 506 splicing modules (x-axis) that were quantified in at least 80% of patients, sorted by the most variant PSI for each module (y-axis). The red cutoff line identifies modules with a variance of one SD (105.21) above the mean (54.26). (**A**, right) The 40 splicing modules above the threshold in B replotted according to SD of PSI. The most variable module is indicated in the plot. See Fig. 2 for details. (**B**) Heatmap showing transcript abundance values (Left) and PSI (Right) for each of the 40 highly variable modules. Rows are sorted by module number from rank sorting in panel A, with a list of these 40 modules on the far right. Where more than one variable module is observed within a gene, these are differentiated by a number after their gene name. Columns in both heatmaps are identically ordered and sorted based on PSI of the most variable module (*EZH2* exon 11/12). Patients were then assigned letters consistent throughout this study. Gray square within the splicing PSI heatmap indicates an unquantifiable module in a given sample.

Table 2.1 – Splicing Modules quantified within AML-associated genes across the PENN-AML patient cohort.

Splicing Module*	Effect‡	Description	Literature Reference°
EZH2.1	LOF	Skipping of exon 11 and exon 12 leads to frameshift and PTC in the upstream exon.	No study found
FBXW7	COF	Inclusion of regions in 5'UTR. A significant difference in translational efficiency among these 5′-UTRs variants has been observed by *in vivo* Luciferase reporter assay and Western blot.	Liu et al. 2012
U2AF1.1	LOF	Splicing together of mutually exclusive exons 3 and exon3b leads to PTC	Chang et l 2019
PTPN11	LOF	Inclusion of cryptic exon	No study found
DNMT3A	COF	Alternative first exon and promoter between exon 6 and exon 7. DNMT3A-short, but not DNTM3A-long, induced EMT, which is a primary characteristic of tumor-initiating cells. The DNMT3A long/short isoforms display differential localization along the mouse ES cell genome.	Manzo et al. 2017
BRAF.1	LOF	Alternative last exon with different 3'UTR regions and lengths that are expected to bind different microRNAs	Marranci et al. 2015
CSF3R	LOF	Inclusion of cryptic region expected to cause frameshift. Truncated transcripts for this gene have been found to be increased in some patients with MDS/ AML.	Lawrence et al. 2018
BCOR.1	COF	Alternative 3' ss in exon 8 controls inclusion of around 34 amino acids with three consecutive serine residues.	No study found
GATA2.1	COF	Alternative 5' UTR regions.	No study found

STAG2.1	COF	Alternative 5' UTR regions.	No study found
TET2.1	LOF	Inclusion of cryptic exon 2b between exons 2 and 3 expected to reduce protein expression.	No study found
ABL1	COF	Alternative first exon.	No study found
STAG2.2	COF	Multiple alternative 5' UTR regions including alternative first exon. Linked with STAG2.1.	No study found
BCOR.2	COF	Skipping of exon 5 which is in-frame and removes 17 amino acids and may affect BCL6-binding domain. Splicing bypassing exon 5 and/or alternative splice acceptor usage at exon 8 results in the previously identified BCOR isoforms. Only isoforms a and b contain sequences required for the interaction with the transcriptional regulator AF9	Srinivasan et al. 2003 Kim et al. 2015
EZH2.2	LOF	Cryptic poison exon 9b between exon 9 and exon 10 with a high degree of conservation.	Kim et al. 2015
BCORL1	LOF	Multi-exon skipping leads to frameshift	No study found
FLT3.1	LOF	Cryptic exon 4b between exon 4 and 5 expected causes a frameshift leading to PTC in exon 5	No study found
ETV6	Unknown	Complex cryptic exon inclusion between exon 3, exon 4, and exon 5 with noticeable conservation of select intronic regions.	No study found
EZH2.3	LOF	Cryptic Poison exon 3b between exon 3 and exon 4 with a high degree of conservation.	No study found
FLT3.2	LOF	Inclusion of cryptic exon 17b between exons 17 and 18 inserts a PTC.	No study found

ASXL1.1	COF	Complex last exon between exons 4 and exon 5 representative of long/short isoform switching.	No study found
SMC1A	COF	Multiexon skipping expected to remain in-frame and change function of translated peptide	No study found
ZRSR2.1	COF	Alternative last exon representative of a switch to a short isoform.	No study found
BRAF.2	COF	Linked with BRAF.1 splicing module	No study found
TET2.2	Unknown	Linked with TET2.1 splicing module	No study found
ZRSR2.2	LOF	Skipping of exon 9 which leads to a frameshift in the upstream exon	No study found
U2AF1.2	COF	Skipping of exon 5 which is in frame and removes 33 amino acids	No study found
BRAF.3	LOF	Inclusion of cryptic exon causes a frameshift PTC expected to result in unproductive transcript.	No study found
NOTCH1	Unknown	Complex first cryptic exon within long intron with noticeable conservation regions	No study found.
EZH2.4	LOF	Linked with EZH2.3 splicing module	No study found.
KIT	Unknown	Inclusion of cryptic exon 20b between exon 20 and exon 21	No study found
TBL1XR1	Unknown	Complex first exon within long intron with multiple conserved regions.	No study found
TET2.3	LOF	Skipping of exon 6 which leads to a frameshift that induces a PTC	No study found.
JAK2	LOF	Inclusion of cryptic exon 16b between exon 16 and 17 which leads to frameshift that creates a PTC	No study found.
ASXL1.2	COF	Skipping of exon 8 in frame and removes part of the ASXH domain	No study found.
RUNX1.1	COF	Skipping off exon 2 removes 13 amino acid sequence	No study found.

MYD88	COF	Skipping of exon 2 removes N-terminal of the peptide and exerts a dominant-negative effect on LPS-induced, TLR4-mediated signaling pathways and inhibits inflammatory cytokine production.	Janssens et al 2002 Lesly & Alper 2013
RUNX1.2	COF	Complex last exon representative of long/short isoform switching. Alternative and premature C-terminus. Runx1 exon 6–related alternative splicing isoforms differentially regulate hematopoiesis in mice.	Tanaka et al. 1995

** Splicing modules are sorted by decreasing variance. Red refers to the 23 highly correlated splicing modules and clustering association discussed in Figure 2.4*

‡ Effects of splicing modules are predicted to be loss-of-function (LOF) or change-of-function (COF) through manual interpretation and literature reference

º Reference studies found elucidating the particular splicing module or a similar event

Splicing co-regulation pattern across AML cohorts. To determine if the splicing variability that I have observed is a unique feature of the PENN-AML dataset, or representative of a pattern within AML, I next queried these 40 highly variable splicing modules within independent sample cohorts, specifically the large and publicly available BeatAML patient cohort [95] as well as a cohort of healthy myeloid cells [119]. For this, I analyzed data from ~440 AML patient samples previously published as part of the BEAT-AML project. Overall, I found that splicing of the AML-associated genes is generally more variable in both the PENN and BEAT-AML cohorts than is observed in normal CD34+ cells (**Figure 2.3A**). Moreover, at least 30 of the 40 modules that exhibited a variance of greater than 160 in the PENN cohort, also had a variance above this threshold in the BEAT-AML samples (**Figure 2.3B**). By contrast, although there is some intrinsic variability in the same splicing modules in CD34+ cells from 17 healthy donors, my analysis revealed much less variation. Therefore, the highly variable splicing events I detect in our in-house AML cohort population is reflective of the AML population at large.

Having found this signature of pre-mRNA splicing variations in the independent cohort of AML patients, I next queried the data for particularly interesting correlations. Alternative splicing is a tightly co-regulated process, where one RBP can regulate multiple splicing events. Thus, I reasoned that multiple splicing variations maybe altered in a coordinated fashion within the same AML patient. To explore this possibility, I generated a pairwise matrix of Spearman rank correlations of all the 40 highly variable splicing modules. To identify co-regulated splicing modules in the matrix, I applied a hierarchical clustering method to reorder the splicing modules features by their distinct distance. Indeed, the pairwise-correlation clustering analysis revealed that 23 of these splicing modules (including EZH2.1) show a high degree of correlation between one another in

both the PENN and BEAT-AML cohorts, suggesting that these events are co-regulated with one another (**Figure 2.4A**). Importantly, this cluster of co-regulation between splicing modules is not observed in the normal CD34+ cells. This means that although I found significantly lower yet intrinsic variability present in healthy myeloid cells, the co-regulated aspect of these splicing modules is limited to AML patients. Therefore, the co-regulated pattern that I observe in AML patients may be arising from some unidentified source of oncogenic dysregulation of alternative splicing.

The co-regulatory pattern observed in the data described in Figure 2.3 is indicative of the presence of subclusters of AML patients with abnormal, yet coordinated, pattern of alternative RNA splicing. Followingly, I performed a pairwise comparison of patient samples using only the information from the MAJIQ-PSI quantification of the 40 highly variable splicing modules described above. Indeed, hierarchical clustering of the patient similarity matrix reveals two visually distinct clusters of patients in the large BEAT-AML cohort (**Figure 2.4B**). The sample size of both the PENN-AML and CD34 datasets is inadequate for patient-patient clustering analysis. However, I can observe in the PENN-AML heatmap that there are at least 4 samples that deviate from the rest of the cohort (**Figure 2.4C**). Additionally, as expected, there was no visual indication of uncorrelated samples in the CD34+ dataset. Moreover, the subset of the overall AML patient samples that had increased skipping of *EZH2* exon 11 and 12 (PSI > 25%) significantly overlapped with the smaller of the two distinct cluster patients (**Figure 2.4D**).

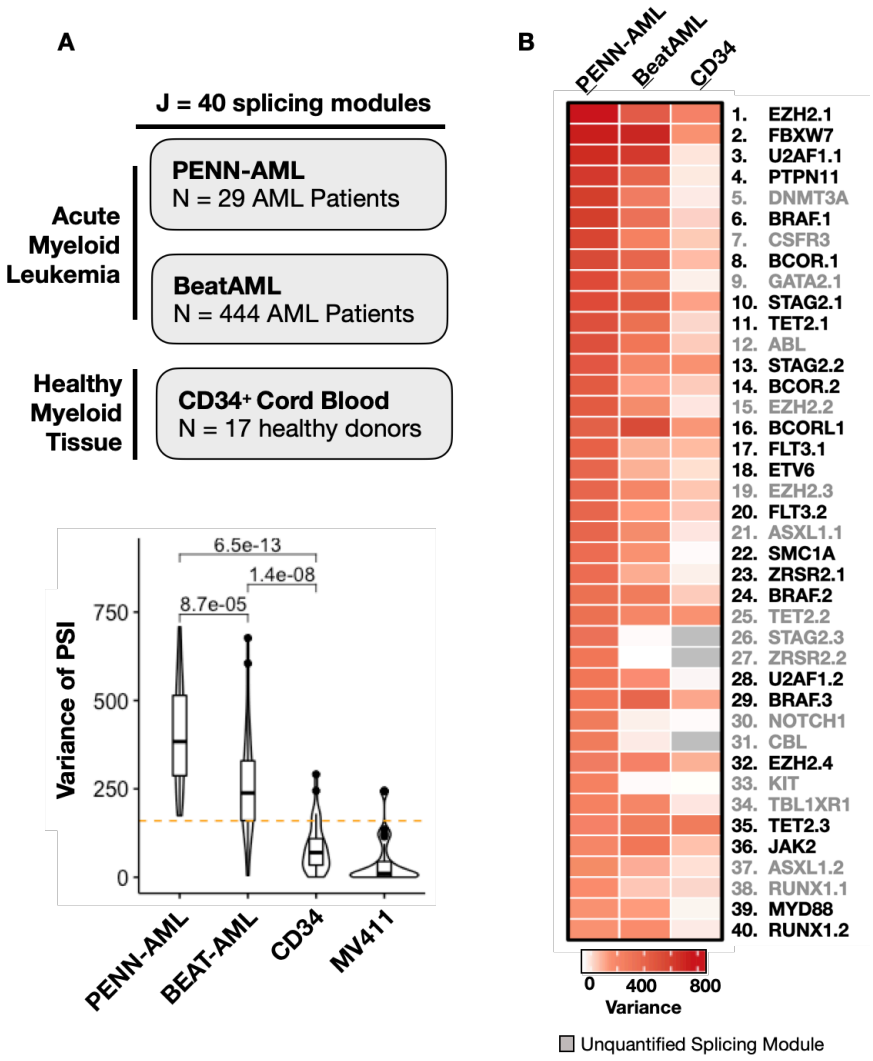


Figure 2.3 –**Highly variable splicing variations are observed in additional AML cohort.** (**A**) Variance of PSI for the 40 highly variable splicing modules quantified in the PENN cohort from Fig. 1, plotted also for that observed in 444 samples from the Beat AML study or 17 healthy CD34+ cell samples from Leucegene publicly available data. The orange dashed line indicates the threshold variability from Figure. 2.2. (**B**) Heatmap of the variance in PSI observed in the Beat AML cohort and normal CD34+ control cells of the 40 most variable modules from the PENN cohort. Modules are named as in Figure 2.4. Colors of module names are based on presence in black or gray clusters in Figure 2.4A.

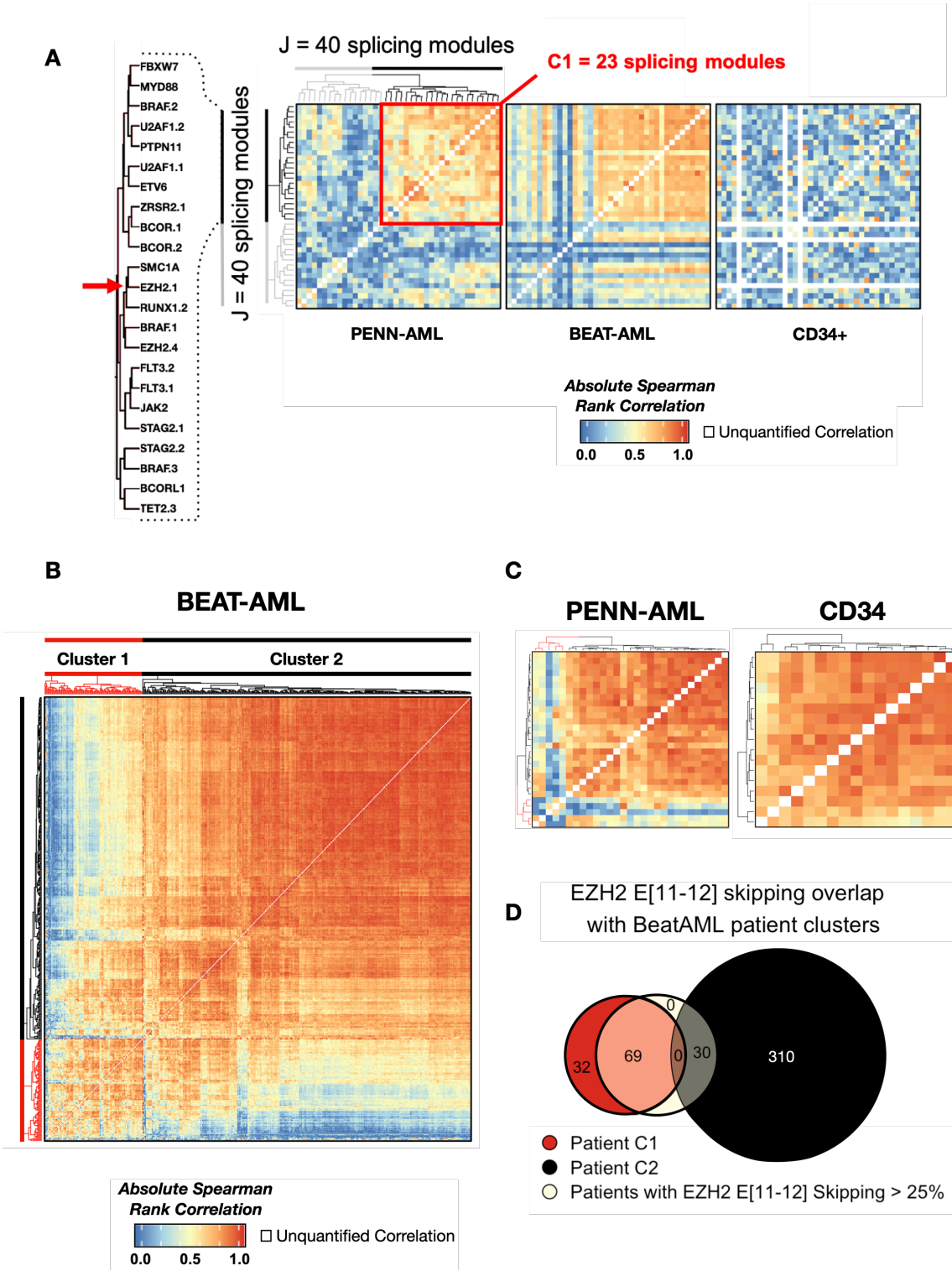

A
J = 40 splicing modules
C1 = 23 splicing modules
J = 40 splicing modules
FBXW7
MYD88
BRAF.2
U2AF1.2
PTPN11
U2AF1.1
ETV6
ZRSR2.1
BCOR.1
BCOR.2
SMC1A
EZH2.1
RUNX1.2
BRAF.1
EZH2.4
FLT3.2
FLT3.1
JAK2
STAG2.1
STAG2.2
BRAF.3
BCORL1
TET2.3
PENN-AML
BEAT-AML
CD34+
Absolute Spearman Rank Correlation
Unquantified Correlation
0.0 0.5 1.0
B
BEAT-AML
Cluster 1
Cluster 2
Absolute Spearman Rank Correlation
Unquantified Correlation
0.0 0.5 1.0
C
PENN-AML
CD34
D
EZH2 E[11-12] skipping overlap with BeatAML patient clusters
32
69
0
0
30
310
Patient C1
Patient C2
Patients with EZH2 E[11-12] Skipping > 25%

Figure 2.4 – **Co-regulation of splicing variations in AML-associated genes.** (**A**) Spearman correlation matrix showing pairwise comparison between each of the variable modules with all others shown as a heatmap. Hierarchical clustering revealed 23 highly correlated splicing modules (red box & black bar, names listed on left). Hierarchical clustering was done based on the PENN (left) data and the same ordering revealed strong clustering and correlation in the Beat AML data (Middle) but not in the CD34+ cells (Right). (**B**) Clustering of BEAT-AML patient-to-patient pairwise comparison. Spearman correlation was used to generate the patient similarity matrix (**C**) Patient pairwise comparison and clustering of PENN-AML (left) and CD34 normal donors(right).

Splicing variations in *EZH2* disrupt functional protein expression. The goal of this dissertation work is to understand how splicing variability contributes to protein expression and function in patients. Thus, I initially focused on the gene *EZH2* because it contains the most highly variable splicing module and is expected to cause a frameshift that disrupts translation of full-length protein (**Figure 2.5**). The *EZH2* gene encodes the catalytic component of polycomb-repressive-complex-2 (PRC2) that confers di- and trimethylation on lysine 27 of histone H3 (H3K27me2/3) [88]. Specifically, skipping of exon 11 and/or 12 in the *EZH2* gene (Fig. 2A, splicing module 1) results in either case in a premature termination codon (PTC) prior to the catalytic SET domain. Skipping of these exons is not observed in two leukemia-related cell lines, MV411 and HL-60, and only minimally in normal CD34+ cells, but varies from 99% (e.g., patient CC) to 0% (patient A) in the PENN cohort (**Figure 2.5C**). This complex exon skipping event in *EZH2* has not previously been described, despite its prevalence in the PENN-AML and BEAT-AML data, presumably because unlike the analysis done here, most splicing quantification methods focus on previously annotated and simple binary splicing choices and exclude events with more than two outcomes. Additionally, I detect a second splicing module in *EZH2* that also introduces a PTC, in this case through inclusion of variable exon 9b. Inclusion of

EZH2 exon 9b has previously been observed in some AML and MDS patients and has been shown to induce nonsense-mediated decay (NMD) which in turn leads to loss of protein expression and a corresponding reduction in H3K27me3 levels [92]. Consistently, I observe over 40% inclusion in exon 9b in six of the PENN patients (**Figure. 2.5C** bottom, patients M, P, R-T, V). However, there is no correlation between inclusion of exon 9b and skipping of exons 11/12 in EZH2. Thus, both of these splicing modules within *EZH2* represent alternative avenues through which this gene may be functionally downregulated in AML blasts.

Importantly, radiolabeled RT-PCR performed by Michael J. Mallory of the Lynch Lab validated the occurrence of *EZH2* exons 11/12 skipping in several of our AML patient cells and we observe the predicted decrease in EZH2 protein in these same samples relative to patients that exhibit little-to-no exon 11/12 skipping (**Figure 2.6A**). Although high EZH2 protein in MV411 relative to patients corresponds to high *EZH2* RPKM, I am still able to observe a 10-fold decrease in EZH2 protein in patients with high exon 11/12 skipping (AA and CC) relative to AML patients with high inclusion (N and W). Heterozygous loss-of-function (LOF) somatic mutations in *EZH2* have been described in AML but are only observed in a small percentage of patients. Consistently, I observe that 4 of our 29 patients (14%) harbor a somatic mutation in EZH2 (**Figure 2.5D**). By contrast, I observe 8 additional patients (28%) for which greater than 50% of their EZH2 transcripts fail to encode full-length protein due to skipping of exon 11/12 and/or inclusion of exon 9b (patients M, R-T, V, AA-CC). Therefore, taking splicing patterns into account in the transcriptomic analysis triples the percentage of EZH2-deficient patients (14% to 42%), consistent with the notion that splicing variation is an underappreciated contributor to protein dysregulation in AML.

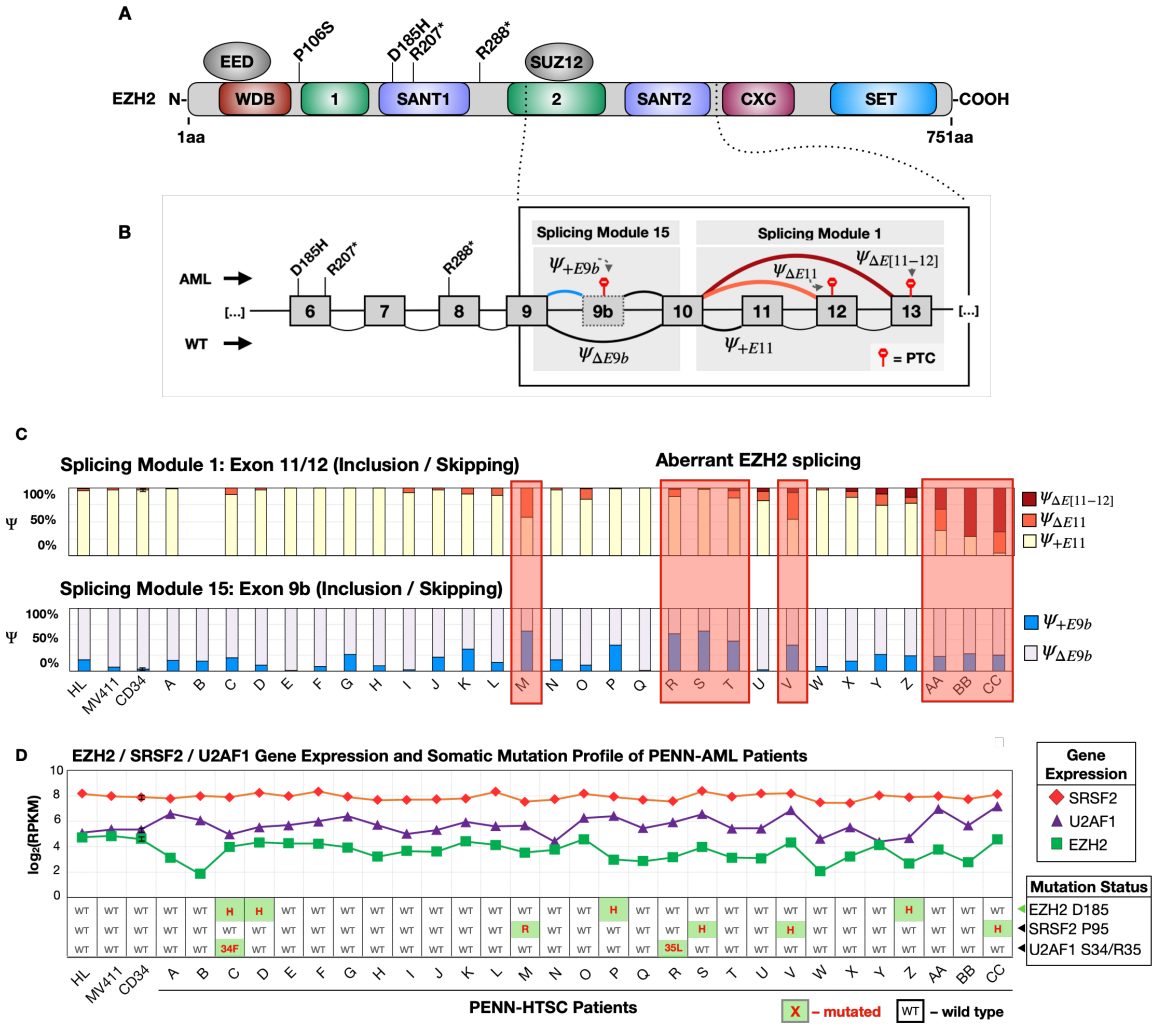


Figure 2.5 – **Splicing, expression, and mutational analysis of *EZH2* across the PENN-AML patient cohort.** (**A**) Schematic diagram of EZH2 protein domain structure and (**B**) the corresponding exon-junction connectivity of the *EZH2* gene and with highlighting of the two most variable splicing modules in *EZH2*. Also indicated are known AML- related *cis* mutations in *EZH2* found in the PENN cohort (asterisk indicates nonsense mutations) and induced premature stop codons (PTC, red). (**C**, top) PSI values for exon 11 and 12 skipping and (**C**, bottom) PSI values for exon 9b inclusion in patients, cell lines (HL-60 and MV411), and CD34+ normal cells. NQ for patient B indicates this LSV was not quantifiable in this patient. Red squares indicate patients with *EZH2* splicing variations at PSI levels that are expected to functionally downregulate the resulting protein (**D**) Gene expression (Top) and mutational status (Bottom) of EZH2 as well as two known regulatory proteins SRSF2 and U2AF1 in the samples corresponding to B and C above.

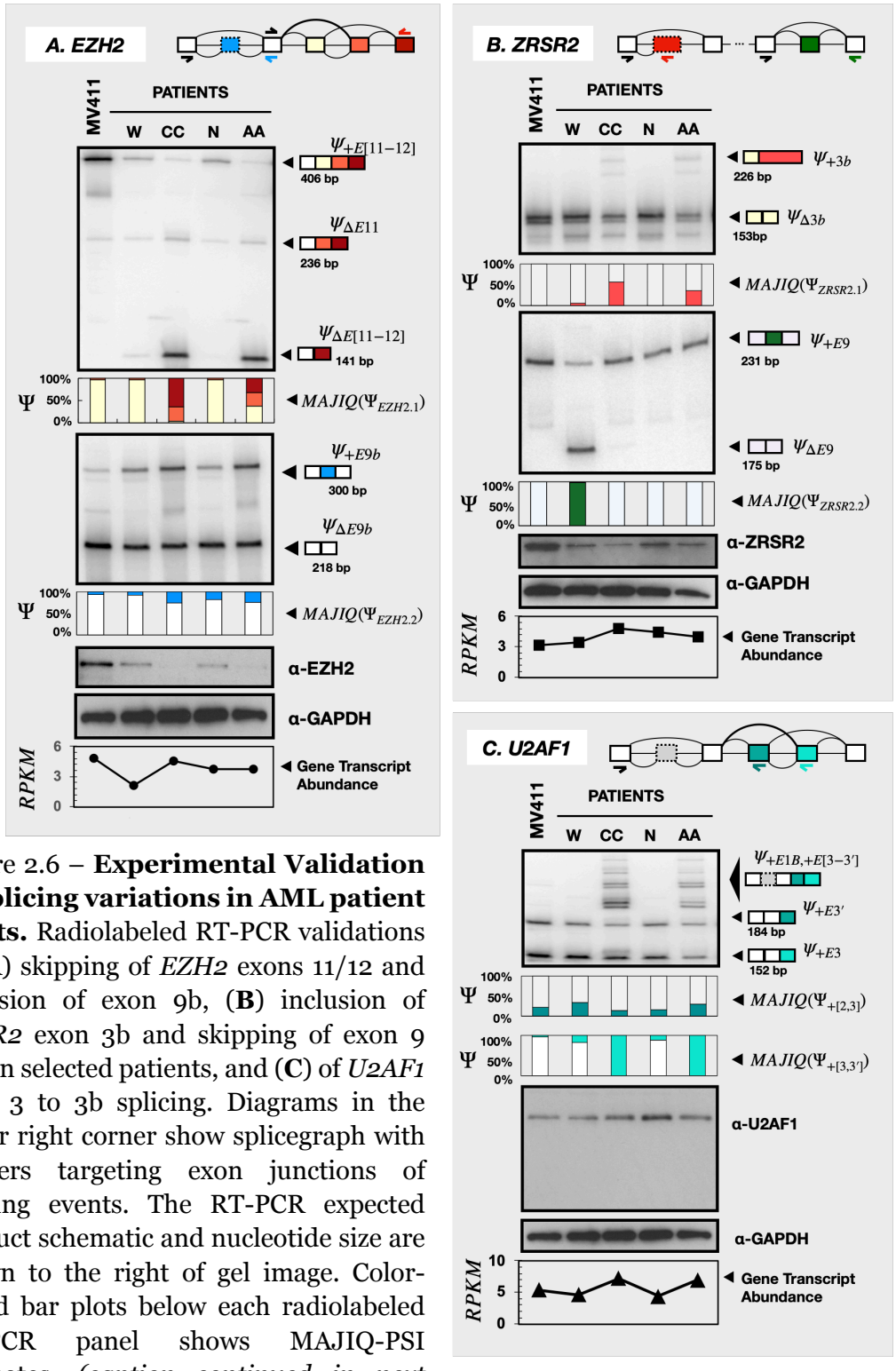


Figure 2.6 – **Experimental Validation of splicing variations in AML patient blasts.** Radiolabeled RT-PCR validations of (**A**) skipping of *EZH2* exons 11/12 and inclusion of exon 9b, (**B**) inclusion of *ZRSR2* exon 3b and skipping of exon 9 within selected patients, and (**C**) of *U2AF1* exon 3 to 3b splicing. Diagrams in the upper right corner show splicegraph with primers targeting exon junctions of splicing events. The RT-PCR expected product schematic and nucleotide size are shown to the right of gel image. Color-coded bar plots below each radiolabeled RT-PCR panel shows MAJIQ-PSI estimates. *(caption continued in next*

(Figure 2.6 continued). The gel images below radiolabeled RT-PCR panels show western blot analyses. Below each blot are plots of log2(*RPKM*) transcript expression levels pertaining to each particular gene panel. (radiolabeled RT-PCR and western blots were performed by Michael J. Mallory from the Lynch Lab).

Previous studies have shown that one driver of *EZH2* exon 9b inclusion is mutation of the splicing regulatory protein SRSF2. Specifically, mutation of SRSF2 proline 95 (P95) to histidine, lysine or arginine alters the mRNA binding specificity of this splicing factor, resulting in its binding to sequences within EZH2 exon 9b to promote exon 9b inclusion [92]. Using the RNA-Seq short-read data, I determined that, of the six patients in the PENN cohort that exhibit highest inclusion of EZH2 exon 9b, two carry the P95H mutation in SRSF2, while a third has the P95R mutation (**Figure 2.5D**). A second mutation that has been shown to correlate with high inclusion of *EZH2* exon 9b in MDS is a S34F mutation in the core splicing factor U2AF1 [120]. I detect this mutation in patient C, which has low inclusion of exon 9b, while mutation in the neighboring residue in U2AF1(R35L) co-exists with high exon 9b inclusion in patient R. The U2AF1 R35L mutation has been detected in AML but has not previously been shown to drive inclusion of EZH2 exon 9b. Notably, at least one instance of high inclusion of *EZH2* exon 9b lacks known mutations in SRSF2 or U2AF1 (patient T) while another patient (CC) has the SRSF2 P95H mutation but low inclusion of exon 9b, suggesting other yet-unidentified drivers of this splicing event. Moreover, no mutations in SRSF2, U2AF1 or EZH2 correlate with skipping of *EZH2* exons 11/12, nor does this splicing event correlate with the gene expression level of *EZH2*, *SRSF2* or *U2AF1* (**Figure 2.5D**). Therefore, none of the features previously described as correlated with functional *EZH2* expression in AML predict the loss of respective protein observed upon skipping of exons 11 and/or 12.

Splicing variations in Splicing Factors. Among the genes that contain the co-regulated splicing events are those that encode U2AF1 and ZRSR2, both splicing factors themselves that participate in the recognition and use of 3' splice sites [121; 100]. The most highly variable splicing module in *U2AF1* involves the inclusion of both exon 3 and 3' which results in a stop codon in exon 3' (**Figure 2.7A-B**). The U2AF1 gene contains duplicated tandem exons between exon 2 and exon 4. These two duplicated tandem exons (3a and 3b) are mutually exclusive in splicing and yield two highly similar protein isoforms, U2AF1a and U2AF1b. They are evolutionary conserved and only differ by seven amino acids in the final protein products (97.1% identity) [122]. Although radiolabeled RT-PCR confirmed the inclusion of both exon 3 and 3' by, this event represents less than 25% of the transcripts and thus, not surprisingly, does not correlate with a detectable decrease in full length U2AF1 protein (**Figure 2.6C**). I also note the presence of another splicing module (U2AF1.2, module 28) that involves skipping of exon 5, which is predicted to result in an in-frame deletion (**Figure 2.7A, C**). However, I also don't observe any smaller U2AF1 protein in the Western blot. Although there is some heterogeneity in the expression of U2AF1 [var($RPKM_{U2AF1}$) = 0.51] across PENN-AML patients, there was no significant correlation between the splicing of *U2AF1* and the RPKM of its own transcripts. The patients that have higher splicing variability in *U2AF1*, seem to have increase transcript expression in terms of RPKM, which may suggest that the patient AML blast is attempting to compensate for the molecular impact of altered *U2AF1* splicing, however, it remains an open question the particular impact these may have in functional protein abundance, and thus, I cannot make any predictive statements about the spliceoforms that I find in *U2AF1*.

By contrast, in the AML patients for which we had protein samples, I do observe a clear reduction of full-length ZRSR2 protein that correlates with splicing variation

(**Figure 2.6B, Figure 2.8A-D**). Four patients (V, AA, BB and CC) exhibit greater than 25% splicing from exon 3 to an alternative terminal exon 3b, rather than to exon 4 and beyond to encode the full-length protein (**Figure 2.8A** top, and **2.8B**) The alternative termination exon 3b is annotated to have its own 3'UTR, so the expression of a smaller protein isoform is likely. However, I do not observe the presence of a smaller isoform by Western blot analysis. I also detect near-complete skipping of ZRSR2 exon 9 in one patient (patient W), which introduces a premature stop codon (**Figure 2.6B, 2.8A** bottom, and **2.8C**). By Western blot, there is slightly less ZRZR2 protein in patient W compared to patient N, which seems has the highest levels protein of the 4 patients tested. Patients CC and AA have lower protein expression which is correlated with the inclusion of exon 3b. Similar to all of the splicing events that I have discussed, the overall transcript expression for *ZRSR2* is uncorrelated with *cis*-acting splicing variations. Additionally, there slightly less heterogeneity in the transcript expression of *ZRSR2* across the AML patient samples [var($RPKM_{ZRSR2}$) = 0.34] compared to *U2AF1* and *EZH2* [var($RPKM_{EZH2}$) = 0.58] (**Figure 2.8D**). Therefore, the identification of these two splicing events in *ZRSR2* uncovers multiple avenues by which the encoded protein may be functionally downregulated. Overall, the identification of splicing events within splicing factors themselves that are part of the co-regulatory cluster generates new questions with regards to how the splicing modules that I have uncovered are manifested in AML blasts (See Discussion).

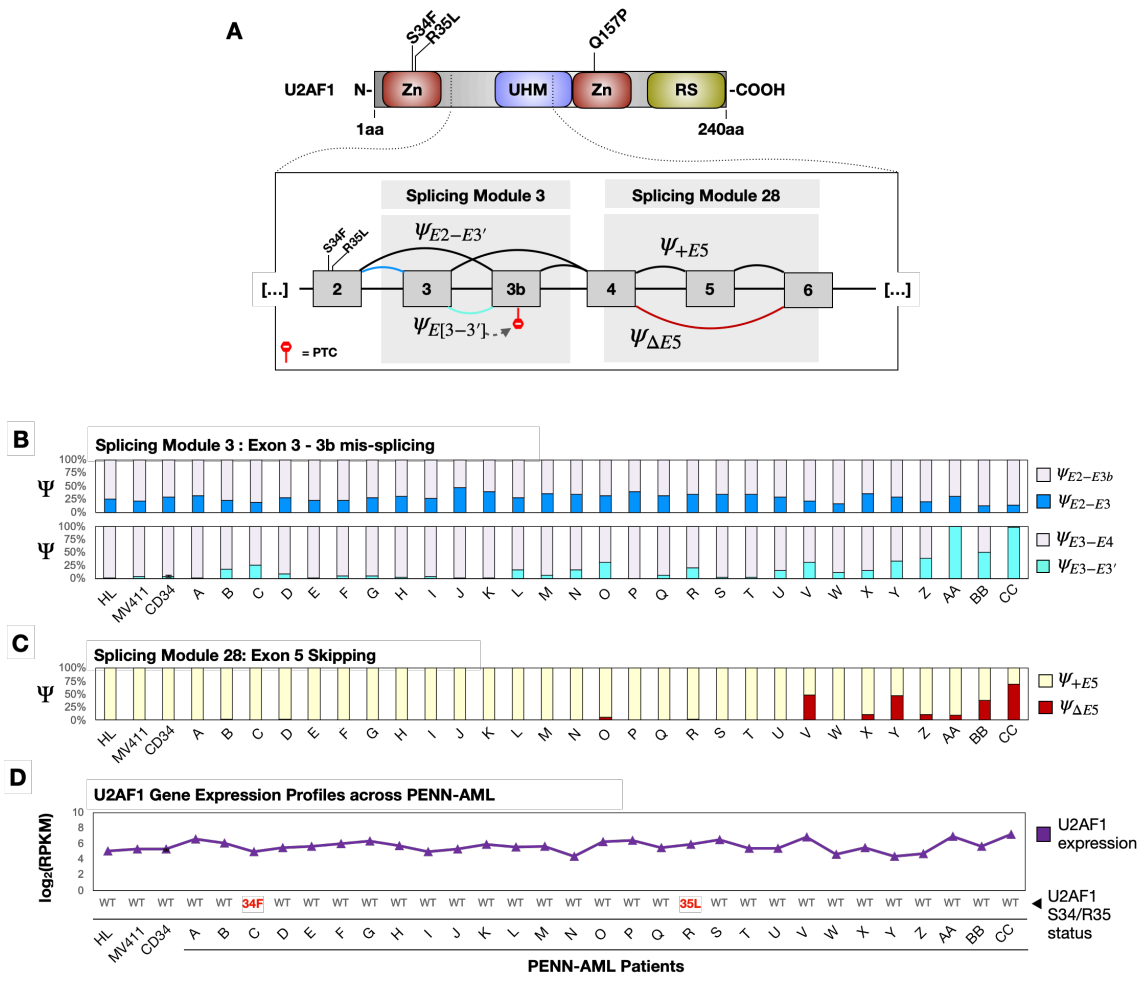


Figure 2.7 – **Splicing, expression, and mutational analysis of *U2AF1* across the PENN-AML patient cohort.** (**A**) Schematic diagram of the U2AF1 protein domain structure and the corresponding exon connectivity in *U2AF1* gene with highlighting of modules 3 and 28 from Figure 1 and 3. (**B**) PSI values for module 3; exon 2 splicing to alternate exons 3 or 3b or 4 (top) or 3 to 3b (bottom) in PENN cohort patients, cell lines (HL-60, MV411) and CD34+ normal cells. (**C**) PSI values for exon 5 skipping (module 28) in PENN cohort patients, cell lines (HL-60, MV411) and CD34+ normal cells. (**D**) Gene expression (top) and mutational status (bottom) of U2AF1 in the samples corresponding to panel B and E.

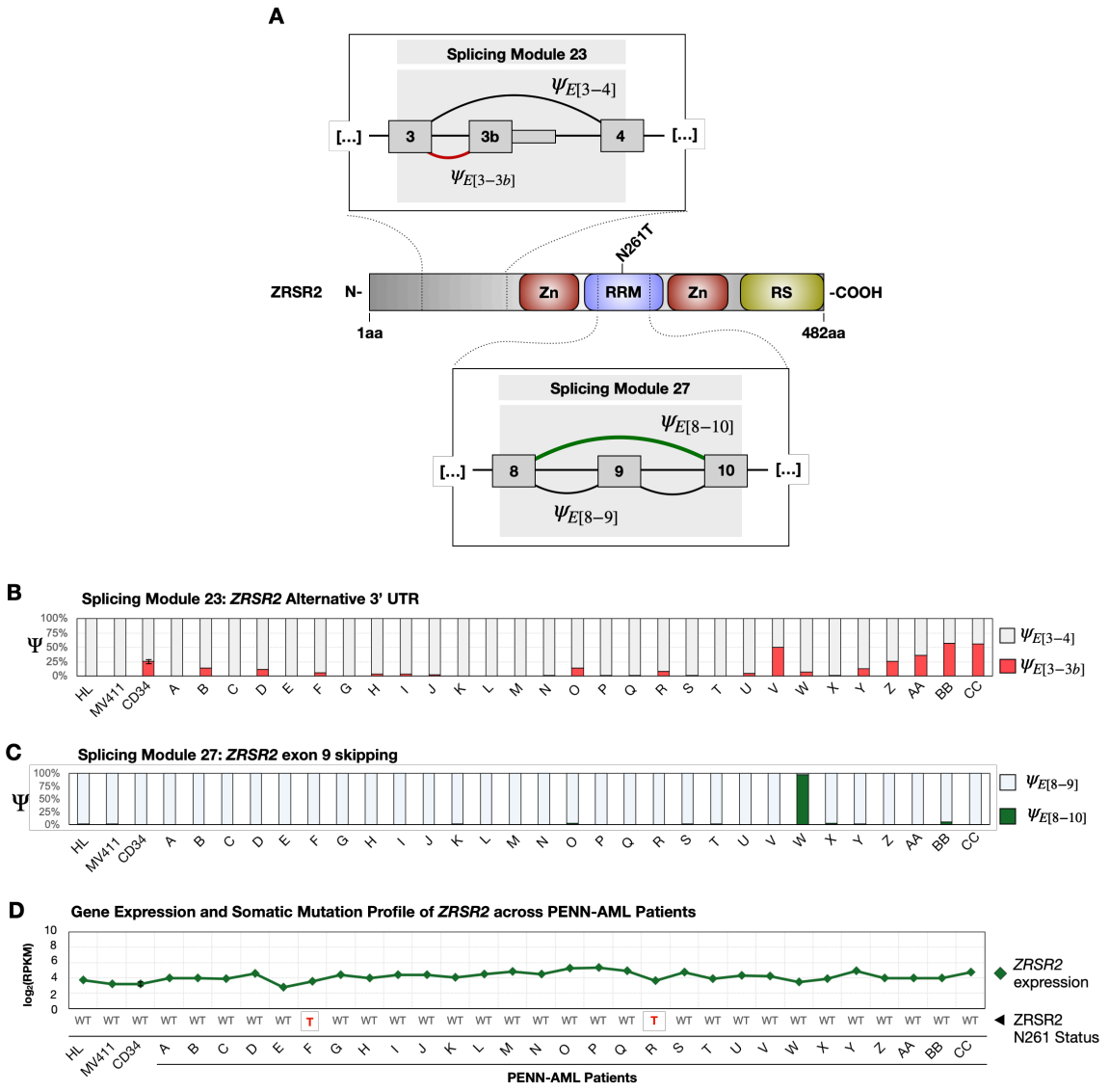


Figure 2.8 – **Splicing, expression, and mutational analysis of *ZRSR2* across the PENN-AML patient cohort.** Splicing, expression and mutational analysis of ZRSR2 across the PENN AML patient cohort. (**A**) Schematic of ZRSR2 protein domain structure, and exon connectivity in ZRSR2.1, module 23 from Figure 1 and 3. and exon connectivity in ZRSR2.2 (module 27, exon 9 skipping). (**B**) PSI values for exon 3 splicing to alternate exons 3b or 4 in PENN cohort patients, cell lines (HL-60, MV411) and CD34+ normal cells. (**D**) PSI values for exon 9 skipping in PENN cohort patients, cell lines (HL-60, MV411) and CD34+ normal cells. (**E**) Gene expression (top) and mutational status (bottom) of ZRSR2 in the samples corresponding to panel B.

Splicing Regulatory Elements associated with co-regulated splicing modules. Motif analysis for the highly variable modules revealed enrichment of a purine-rich element in the exons and a U-rich element in introns (T-rich in the genomic signature) as compared to non-variable modules (**Figure 2.9A**). ZRSR2 does not have a defined binding motif, but it is possible that the U-rich motif reflects sequences in a 3' splice site that confer regulation by ZRSR2. Accordingly, I also determined if expression of known purine-rich or U-rich binding proteins exhibited any correlation with the splicing of the highly variable modules. Interestingly, I there was stronger correlation between gene expression of *SRSF1* and most of the 23 co-regulated modules from cluster C1 discussed in Figure 3 [$\mu_{corr(SRSF1,\ C1)}$ = 0.28] compared to the rest of the uncorrelated 17 modules [$\mu_{corr(SRSF1,\ C2)}$ = 0.0981] (**Figure 2.9B**). A similar but weaker trend was also observed for hnRNP F, but not for any of several other genes that encode proteins that bind purine-rich (hnRNP H, hnRNP A1, TRA2A, SRSF7) or U- rich (TDP43, TIA1) motifs, or other splicing regulatory factors that are mutated in some AML patients (U2AF1, SRSF2).

Although *SRSF1* was not included as an AML-associated gene, it is known that poison exon 4 regulates active expression of this particular gene. I was able to find increased inclusion of *SRSF1* poison exon 4 in the patients that exhibit altered splicing of *EZH2*, *ZRSR2*, *U2AF1* and the other splicing variations in the co-regulated cluster, suggesting that this exon could be part of the co-regulatory splicing network that I have uncovered. The functional downregulation of SRSF1 protein expression via poison exon 4 inclusion could be influencing the splicing of other splicing modules. Michael Mallory from the Lynch Lab performed shRNA transfection in Mv411 cells and successfully knocked down SRSF1 protein, however, we did not observe changes in inclusion across splicing modules within genes of interest (*EZH2*, *BRAF*, *BCOR* or *U2AF1*). Thus, altered

splicing and expression of *SRSF1* may be a downstream effect of the aforementioned co-regulatory feedback loop instead of a global driver of such variations.

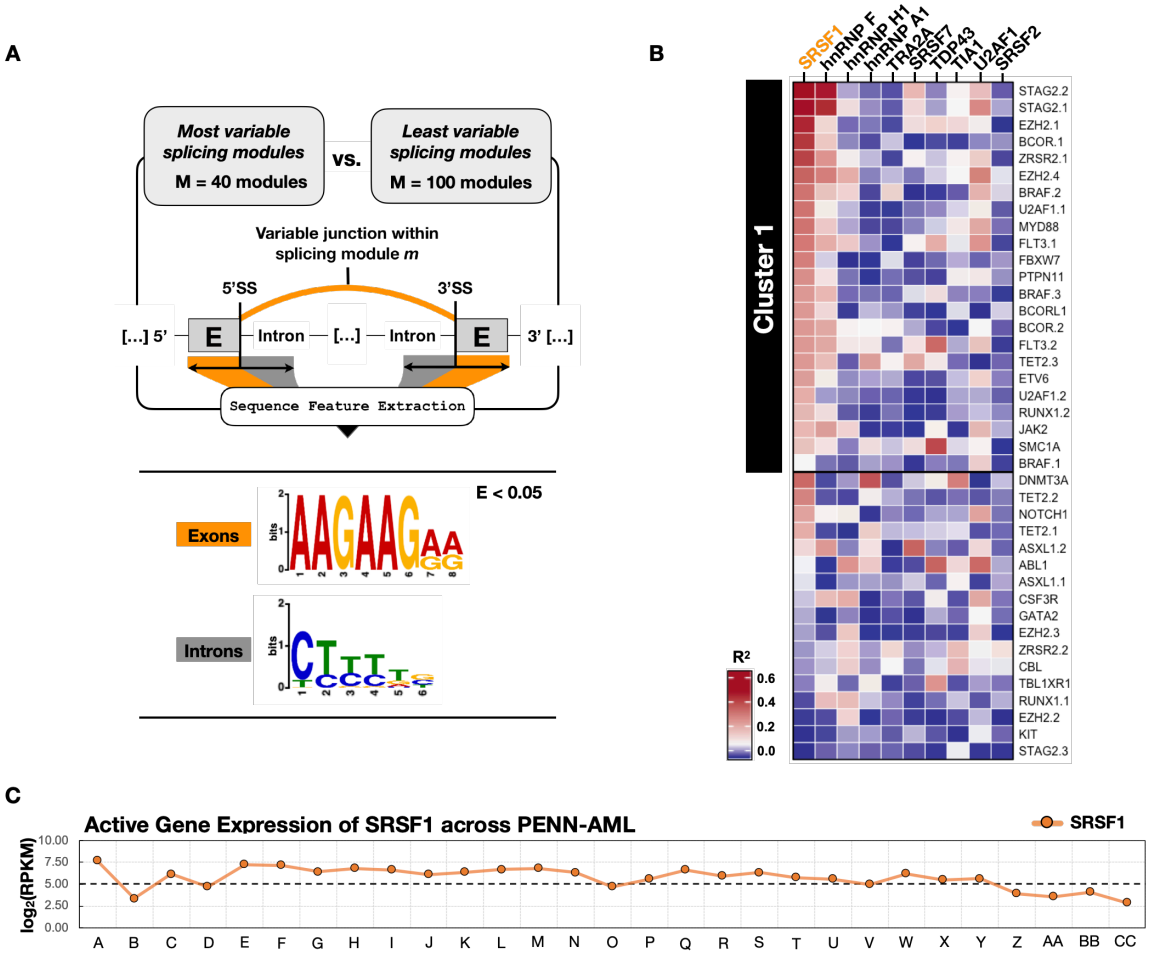


Figure 2.9 – **Poly-AAG motifs surrounding variable exons suggest mechanisms of regulation**. (**A**) Motif enrichment analysis of exonic and intronic sequences encompassed in the high variable splicing modules. (**B**) Heatmap of correlation between expression of RBPs known to bind to GA- or U-rich sequences and PSI of highly variable splicing modules. (**C**) *SRSF1* active gene expression levels across the PENN-AML patient cohort. *SRSF1* active gene expression was calculated by multiplying the gene abundance values [log2(RPKM)] with the levels of inclusion (PSI) of known poison exon 4.

Discussion

In sum, the analyses presented in Chapter 2 highlight alternative RNA splicing as a major contributor to gene dysregulation in AML. In particular, I find that alternative splicing of *EZH2* and *ZRSR2* lead to a higher penetrance of loss of function molecular lesions of these genes in AML than has been previously recognized. While I focused this study on splicing variations that reduce protein expression and function by either introducing PTCs or frameshifts, several of the highly variable splicing modules I detect in the AML patients cause internal *in-frame* changes to the open reading frame as well as alter the N-terminus or C-terminus of the encoded protein by including alternative first or last exons respectively. (**Table 2.1**). Of note, for over half of the genes that harbor gain- or change-of function genetic mutations in AML as well as highly variable splicing modules, the respective altered splicing is predicted to change the function of the protein rather than introduce a PTC. This consistency in the impact of molecular alterations within the same gene via two distinct avenues suggests that these in-frame splicing variations may also phenocopy genetic mutations as I have observed for the LOF mutations. However, more work will be required to fully elaborate the functional consequences of these in-frame splicing variations and compare their effect to genetic mutations in the same genes (Discussed in Chapter 4).

Questions still exists about the mechanism that drives the observed splicing variability. The lack of correlation between splicing patterns and the mutation status or expression of the same gene argues against cis-regulatory effects. By contrast, the fact that several of the splicing variations exhibit strong co-correlation across the AML population implies the dysregulation of a common upstream trans-acting factor(s) that controls many or all of these events. One mechanism I explored was whether altered splicing of *U2AF1*

or *ZRSR2*, both encoding factors involved in 3' splice site selection, leads to altered splicing of the co-occurring modules. Similar to the experiments conducted for *SRSF1,* shRNA-mediated knockdowns of *ZRSR2* or *hnRNP F* in MV411 cells also had no impact on splicing of the correlated modules in *EZH2*, *BRAF*, *BCOR* or *U2AF1*. Thus, I was unable to confirm this prediction as depletion of U2AF1 or ZRSR2 in MV411 cells had no impact on other splicing events tested. It might be that MV411 cells do not recapitulate the splicing context of AML cells. Consistently, I observed that MV411 cell lines did not harbor alternative splice variants when performing the experimental validations of splicing modules showed in Figures 2.5 to 2.8. This suggests that although the MV411 cell line is commonly used as a model of AML, these cells have a high fidelity in mRNA splicing which may not be easily perturbed via transient transfections. Finally, although the experiments in cell lines were inconclusive, the data does point to potential regulatory factors that may drive at least some of the variability in splicing within the context of AML.

Another approach I took to identify potential trans-activators of these AML splicing patterns was to identify motifs that are in common amongst the co-regulated splicing events. This analysis revealed the enrichment of AAG-repeats and U-rich motifs in the exons and introns, respectively, of regulated modules. The presence of AAG-repeats, in particular, is of interest as I also observed a strong correlation between splicing module variability and expression of *SRSF1*, a protein known to bind purine-rich sequences and to promote metastasis (Refer to Chapter 1). Again, while single knock-down of SRSF1 in MV411 cells did not alter splicing of variable modules in *EZH2*, *ZRSR2* or any other genes tested, I cannot rule out an underlying cooperative effect as well as cell-type specific impact on splicing variability within AML cells. I also note that while ZRSR2 has been implicated in the splicing of U12-type introns, only five of the genes I surveyed have a U12-

type intron (*UPF1*, *BRAF*, *MYH11*, *PTEN*, *SNC3*), and none of these U12 introns were variable across our dataset. Therefore, while I cannot conclusively define a mechanism for the co-regulated splicing observed across AML cohorts, my results aim to motivate and direct future studies of RBPs, to elucidate their potential role in driving the variability in splicing across AML patients.

CHAPTER III

Alternative splicing defines new paths to altered gene function in AML

Introduction

The analysis presented in Chapter 2 was focused on 70 genes that have been widely reported to harbor mutations in sporadic AML; however, splicing mis-regulation could lead to altered function of genes not previously linked to the occurrence of AML. Many patients initially exhibiting non-proliferative bone marrow failure syndromes are prone to the development of AML [123]. Some familial MDS/AML patients harbor mutations in genes that are known to be also mutated in sporadic AML/MDS such as *RUNX*, *CEBPA* and *GATA2* [124; 125; 126] Recently, an analysis of the largest series of AML/MDS families assembled to date revealed that several new loci within a set of 20 genes that predispose families to MDS/AML (**Figure 3.1A**) [127]. Mutations in these 20 genes have not previously been described in sporadic cohorts of AML, and no previous study has specifically focused on querying splicing variations within this set of genes. Therefore, I hypothesized that some of these genes may exhibit variable splicing capable of disrupting protein function and phenocopying the deleterious effects of genetic mutation.

In the following body of work, I characterize an undocumented splicing variation within the gene *DHX34* which results in the introduction of a PTC and consequently downregulates the encoded protein. The DHX34 protein is part of the of DExD/H-type RNA-dependent family of ATPases/helicases, which include major players in spliceosomal RNA-RNA rearrangements and RNP remodeling events [96; 128]. Importantly, loss-of-function mutations in DHX34 were identified in families with early on-set MDS/AML [127]. DHX34 itself is also involved in coordinating NMD activity within the cell, suggesting that the splicing variation within this gene is part of an undocumented autoregulatory feedback loop. This finding suggested to me that AML blasts rely on the functional downregulation

of the NMD pathway to support oncogenic molecular profiles. Therefore, as a follow up analysis, I also performed an additional query of highly variable splicing events within genes that encode well-known NMD factors.

The NMD pathway selectively degrades mRNAs harboring PTCs, that if translated can produce truncated proteins that may result in dominant-negative or deleterious gain-of-function effects [129]. The NMD pathway is a widely known quality-control mechanism that is a highly conserved across eukaryotes [130]. Around 30% of disease-associated mutations generate PTCs [131]. However, the NMD pathways also regulates the steady-state levels of around 10% of human physiological mRNAs to maintain appropriate levels of gene expression in response to cellular needs. In mammalian cells, alternative mRNA splicing activity is tightly coordinated with the selectivity of stop codons targeted for NMD and the post-transcriptional control of gene expression [132]. More specifically, the interaction between the spliceosome-deposited exon-junction (EJC) protein complex and certain NMD factors assembled around the upstream stop codon triggers a series of steps that ultimately lead to mRNA decay – hence the term nonsense.

The EJC core primarily consists of four proteins: RBM8A (RNA-binding motif 8A, also known as Y14), MAGOH, eIF4III (eukaryotic initiation factor 4A3), and MLN51 (metastatic lymph node 51. also known as CASC3), all of which participate in NMD [133; 130]. A translating ribosome terminating prematurely at a PTC leaves downstream one or more EJC complexes, which are not removed from the mRNA and subsequently will recruit the NMD machinery. During PTC-mediated NMD degradation process, a complex comprising of the NMD factor proteins Up-frameshift proteins (UPF1, UPF2, UPF3), the SMG proteins (SMG1, SMG8, SMG9), the translation termination factors eRF1 and eRF3,

and more recently identified DHX34 are all assembled around the PTC to form the *surveillance* complex [134; 135; 136]. Although UPF1, UPF2, UPF3 transiently exist as a complex, these show different cellular localization [137; 138; 130]. The RNA helicase DHx34 functions as a scaffold protein to promote ATP-dependent molecular transitions so that SMG1 phosphorylates UPF1 to activate the surveillance complex into a *decay-inducing* complex [136].

Interestingly, there is evidence of feedback-dependent regulation of the NMD process whereby mRNAs coding for most known NMD factors are targeted themselves by the NMD pathway [139]. In particular, long 3' UTRs of *UPF1*, *SMG5*, and *SMG7* mRNAs have been found to be a main NMD-inducing feature of these mRNAs, suggesting that long 3' UTRs might be a frequent trigger of NMD. Presumably, dysregulated alternative mRNA splicing patterns within NMD factor genes are predicted to result in wide-spread changes to the NMD pathway activity, which can be leveraged by the cancer cell to promote oncogenicity. Although the NMD pathway activity has been widely studied, overall, splicing variations within NMD factors genes have been overlooked by previous studies of AML. Therefore, the analysis I present in the following chapter serves to generate new questions about the pathology of AML and its relationship to downregulated NMD.

Results

Splicing mis-regulation in genes associated with familial cases of AML. To test the hypothesis that familial AML genes may undergo alternative splicing in sporadic AML, I repeated the MAJIQ and variability analysis I describe above in Chapter 2. Within the 20 genes associated with germline predisposition of AML, the MAJIQ framework coupled with my custom pipeline identified 186 splicing modules that were quantifiable in at least 80% of the PENN cohort. Indeed, 9 splicing modules in 6 of the 20 genes (**Figure 3.1B**) exhibit variability in the PENN-AML cohort above the threshold set in my initial analysis (**see Figure 2.1**). Similar to what I observed for the other highly variable splicing events in Figure 2.1, variability in splicing of these 9 additional modules is independent of gene expression (**Figure 3.1**) and is also observed in the BEAT-AML cohort **(Figure 3.1D**).

The most variable splicing variation was found within the gene *FANCA* and represents skipping of exon 41 coupled with alternative 3' UTR. In particular, *FANCA* is a large gene (~80 kilobases) and is acknowledged to produce a significant number of different splice variants. Thus, the skipping of exon 41 is expected to potentially change the function of the translated peptide. Of note, although I highlight *FANCA* as a gene associated with germline AML, one study found heterozygotes deletions and reduced expression of *FANCA* in 4 out of 101 cases of sporadic AML. FANCA is part of the Fanconi anemia complementation group, and other members of this pathway include FANCD1 which is also BRCA2, a protein that has been extensively studied for its implication in breast cancer. Lastly, there is crosstalk between the FA pathway and DNA repair pathways, since FA proteins have been found in complexes with DNA repair proteins such as RPA, BRCA1 and ATR [140; 141].

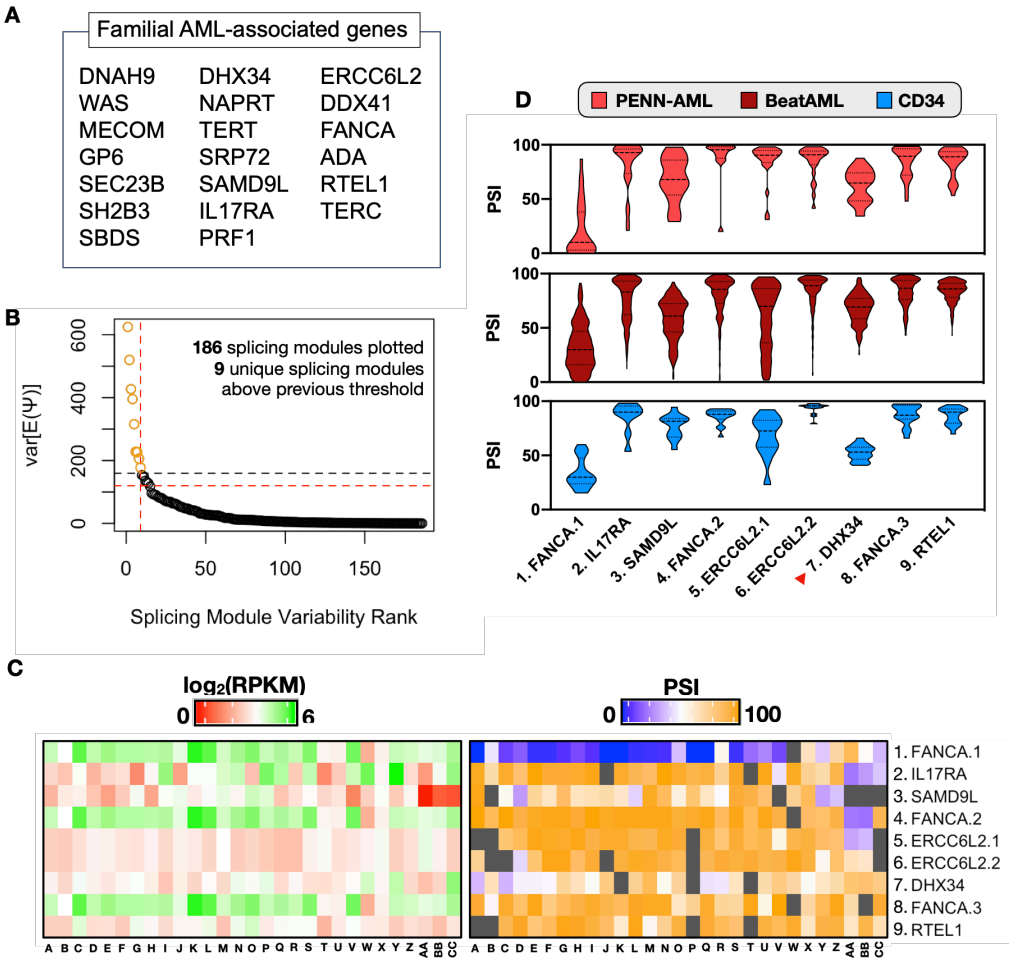


Figure 3.1 – **Splicing of genes newly correlated with familial AML is highly variable across Penn and Beat AML cohorts.** (**A**) Table of familial AML-associated genes queried for splicing variations with MAJIQ-PSI (**B**) Distribution plots as in Fig. 1B of 186 splicing modules (x-axis) from 20 genes linked to familial AML, sorted by the most variant PSI for each module (y-axis). The red cutoff line identifies modules with a variance one SD above the mean, while the black line indicates the variability threshold from the analysis in Figure 2.2. (**C**) Heatmap showing transcript abundance values (Left) and PSI (Right) for each of the nine highly variable modules as described in Fig. 1D. (**D**) Distribution of PSI values for all nine highly variable modules in the PENN (top, light red) and Beat AML (middle, dark red) cohorts, plus 17 CD34+ cells from healthy donors (bottom, blue). Red arrow indicates DHX34 splicing module.

DHX34 poison exon 12b inclusion is significantly higher in AML. While many of the other splicing variations that I quantified within this particular set of germline AML-associated genes may have impact on disease pathology, I find of particular interest the variable module in the gene encoding the RNA helicase and NMD factor *DHX34* (module 7 marked by red arrow in **Figure 3.1D,** and schematic in **Figure 3.2A**). Strikingly, over a third of the Penn and BEAT-AML cohort have greater than 70% inclusion of a 75 bp cryptic exon carrying an in-frame PTC between canonical exons 12 and 13, which I have termed as *exon 12b* (**Figure 3.2B-C**). In contrast, normal CD34 cells overall show a much lower inclusion of exon 12b, centered around 50% ($P_{PENNAML} = 0.026$; $P_{BEATAML} < 0.0001$)

The helicase core of DHX34 as well as other DEAH box proteins is formed by two (RecA)-like domains, a winged-helix domain and a helical bundle domain, known as the Ratchet domain [142] DEAH box proteins also have an accessory C-terminal OB fold (oligonucleotide-oligosaccharide-binding fold domain), which regulates conformational changes in the helicase [143]. *DHX34* exon 12b is inserted into the OB fold coding sequence and would presumably remove the CTD tail that is required to activate NMD. Thus, if the transcript is translated, the resulting truncated protein would be predicted to be non-functional in NMD, with the potential of exerting dominant-negative effects. However, western blots from AML patient blasts that include exon 12b do not reveal the presence of a C-terminally truncated peptide species. Rather, the inclusion of exon 12b correlates with a significant loss of full-length DHX34 protein (**Figure 3.2D-E**). Patients CC and W have higher *DHX34* RPKM than patient N but express less protein, suggesting an attempt to compensate for the lack of functional protein. Patient AA has significantly less *DHX34* RPKM than most other AML patients, suggesting that exon 12b containing transcripts get

degraded. Patient N has less inclusion of exon 12b and thus has significantly higher protein expression. Therefore, inclusion of the premature stop codon in exon 12b appears to destabilize the resulting protein or RNA to decrease expression of active protein, which is consistent with the function of a poison exon.

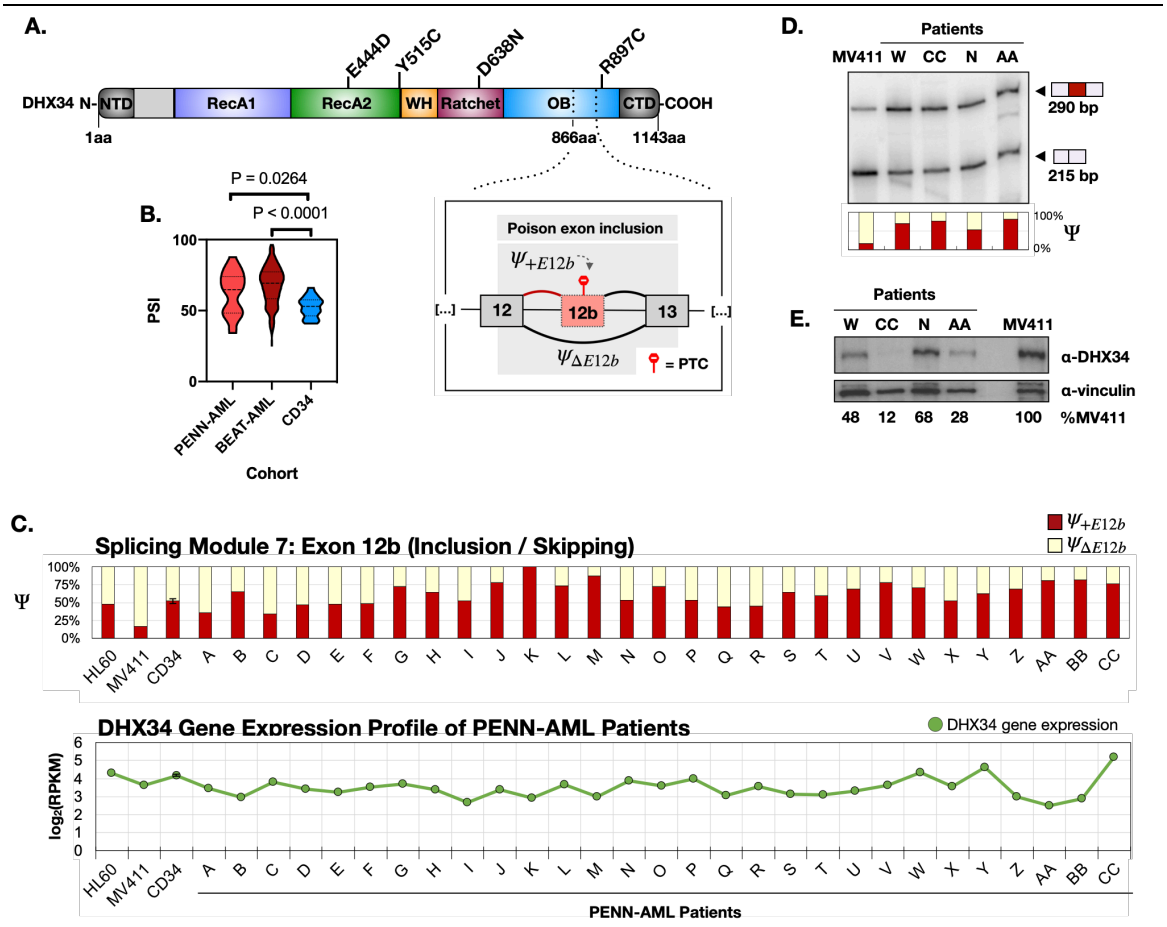


Figure 3.2 – **Inclusion of DHX34 poison exon 12b is significantly increased across both PENN-AML and BeatAML cohorts** (**A**) Schematic of DHX34 protein domain structure and corresponding exon connectivity in the highly variable module. Also indicated are identified LOF mutations in familial AML. (**B**) Comparison of distribution of *DHX34* exon 12b inclusion PSI for the three data sets along with the P value for the differences in distributions. (**C**) PSI values for *DHX34* exon 12b inclusion (Top) and gene expression (Bottom) in PENN patients, cell lines (HL-60 and MV411), and CD34+ normal cells. (**D**) Radiolabeled RT-PCR validation of splicing patterns of DHX34 exon 12b from

selected patients. (**E**) Western blot analysis of DHX34 protein expression from selected patients using an antibody targeting the N terminus of DHX34. (Radiolabeled RT-PCR and Western blot experiment performed by Michael J. Mallory)

***DHX34* poison exon 12b inclusion downregulates RNA decay activity**. Given that the inclusion of *DHX34* poison exon 12b abolishes protein expression, it is expected that this splicing event consequently downregulates NMD activity in the cell. As previously mentioned, NMD selectively degrades mRNAs harboring PTCs and it has been shown to target many transcripts that encode full-length proteins. Importantly, no studies have directly explored target substrates of DHX34 activity. However, one recent study produced a bona fide set of ~1000 target genes of NMD by combining transcriptome profiling of knockdown and rescues of distinct NMD factors, in particular UPF1 as well as two SMG proteins (SMG6/SMG7) [144]. Therefore, I hypothesized that a subset of these well-documented NMD target genes is expressed at higher abundance upon higher inclusion of *DHX34* poison exon 12b inclusion. To address this question, I filtered for around 700 genes that were readily expressed [log2(RPKM) > 1] across the AML cohorts and determined 142 significant correlations between the inclusion of *DHX34* poison exon 12b and the gene expression values of the inferred target geneset. (**Figure 3.3A**). Importantly, I bootstrapped my analysis by generating a random set of 700 expressed genes and calculated the mean number of correlations that I could find by random chance a thousand different times. Strikingly, the set of 700 expressed NMD targets appears to be significantly enriched for genes positively correlating with expression and inclusion of the *DHX34* poison exon 12b across both the PENN and BEAT-AML cohorts (**Figure 3.3B**). In particular, at least four NMD-sensitive genes (*CHD2*, *PARP6*, *METTL22* and *WHAMM*)

that exhibit higher expression in AML patient blasts have previously been linked to important biological functions and the development of cancer [145; 146; 147; 148]

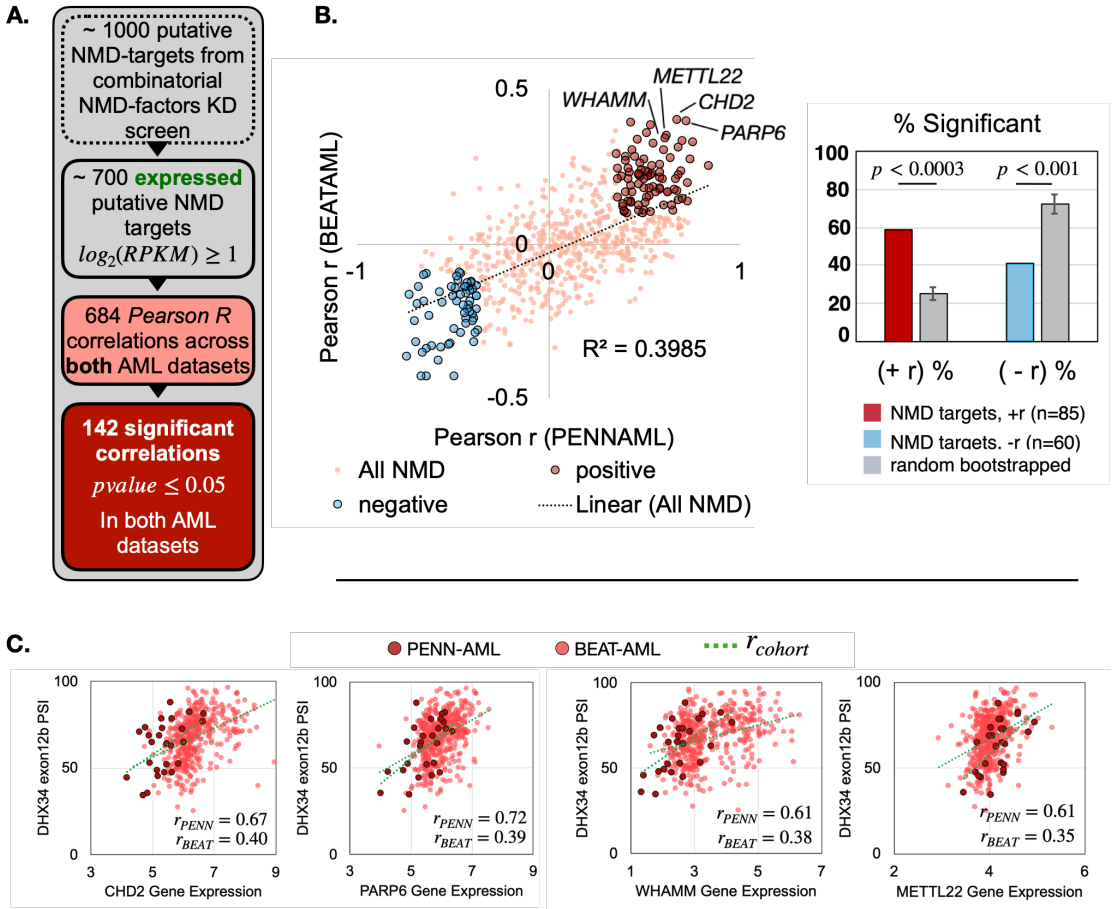


Figure 3.3 – ***DHX34* poison exon 12b inclusion correlates with significantly higher abundance of inferred NMD substrates** (**A**) Pipeline for identifying potential targets of DHX34 NMD activity. (**B**) Plot of the correlation between expression of ~600 documented NMD target genes and DHX34 exon 12b inclusion in PENN (x-axis) or Beat AML (y-axis) patients. Each dot represents a single gene. The observed correlations are in good agreement between the two datasets (Pearson r = 0.3985). Genes for which correlation was significant in both the PENN and Beat cohorts are labeled in red (positive correlation) or blue (negative correlation). On the right is a comparison of the percentage of significantly correlated NMD targets versus the percentage of significantly correlated genes in a random gene set sampled 1,000 times (P value based on a one-sample t test, see Methods). (**C**) Correlation of the expression of specific genes from B with inclusion of DHX34 exon 12b from both the PENN (dark red) and Beat (light red) cohorts.

To more directly test whether the inclusion of exon 12b in *DHX34* leads to deregulation of NMD-target genes, I leveraged the use of a JSL1 (Jurkat Splicing 1) cell line with fused exon 12b and 13. JSL1 cells are a T-cell leukemia cell line heavily used in the Lynch Lab as a model system to study the effects of alternative splicing changes [149] (**Figure 3.4A**). Dr. Rakesh Chatrikhi of the Lynch Lab engineered the exon fusion using CRISPR gene editing system. The fusion from exon 12b 5'ss and exon 13 3'ss was chosen so that the cell is forced to pick the exon 12b 3' ss when splicing the intron between exon 12 and exon 13. Interestingly, we were unable to obtain any clones that had homozygous inclusion of the full 75b exon 12b sequence, because it seems like this higher degree of editing in the cell forces the cell to pick another splice site 50 bp downstream of the original splice site used for the cryptic exon inclusion (Data not shown). However, we were able to obtain clones heterozygous for fused exon12b+13 sequence (HET-12B) that resulted in an increase in the percentage of exon 12b included product by ~20% (**Figure 3.4B**) similar to the difference between the AML patients with average inclusion (50%) and high (75%+) inclusion. Strikingly, this difference in inclusion is sufficient to cause a dramatic decrease in DHX34 protein (**Figure 3.4D**). and a marked increase in the transcript abundance of the NMD gene substrates correlated with *DHX34* exon 12b inclusion in the AML patients (**Figure 3.4E**). Presumably, downregulation of the NMD pathway results from the absence of DHX34-mediated coordination of the decay-inducing complex that forms around the EJC. Therefore, this data strongly suggests that inclusion of *DHX34* poison exon 12b negatively impacts NMD activity within the cell. Furthermore, the increased abundance of non-functional transcripts presumably interferes with various regulatory

functions such as binding of RBPs and microRNAs, as well as may potentially allow translation of truncated proteins that could exert dominant-negative effects.

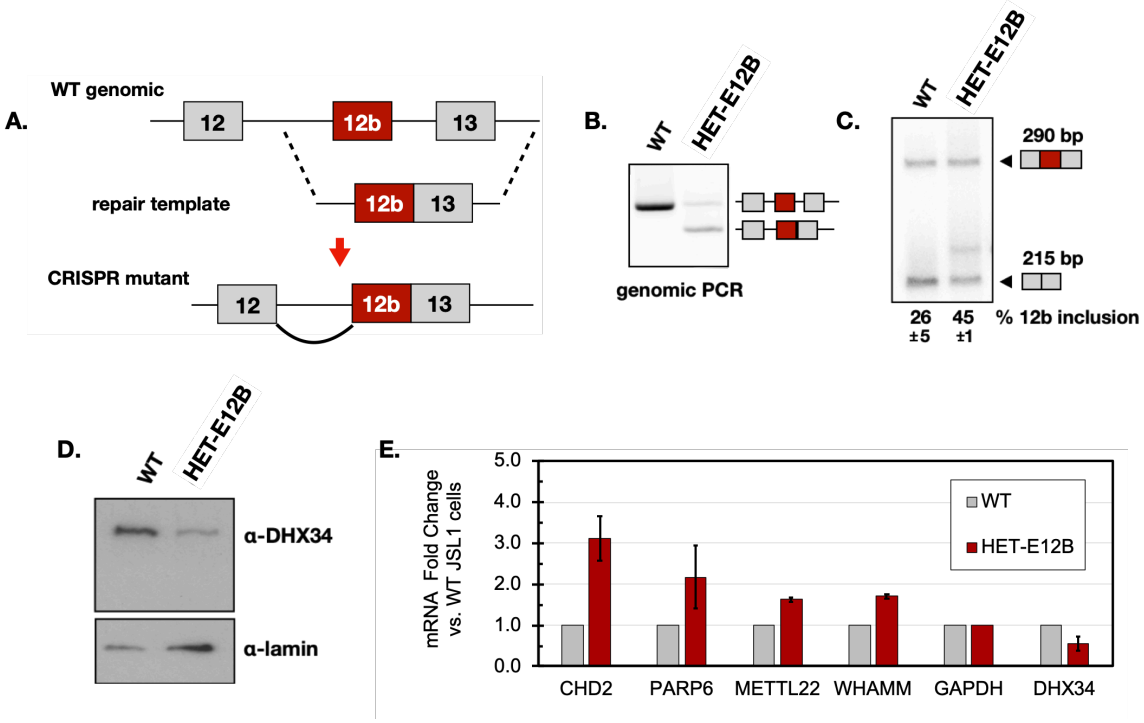


Figure 3.4 –**Forced inclusion of DHX34 exon 12b leads to a reduced DHX34 protein and increased transcript abundance of NMD targets**. (**A**) Schematic of the mutational strategy to force inclusion of DHX34 exon 12b by CRISPR gene editing to fuse exons 12b and 13 (**B**) Genomic PCR analysis of the DHX34 locus in wild-type (WT) cells compared with those containing heterozygous replacement of the exon 12b to exon 13 region with the fused construct. Cell line was engineered, and clone was selected for by Dr. Rakesh Chatrikhi of the Lynch Lab who also confirmed accurate repair by sequencing. (**C**) Radiolabeled RT-PCR analysis of *DHX34* splicing in WT cells compared with the heterozygous mutant from A. Average quantification of inclusion of exon 12b in final total DHX34 mRNA is given (n > 3). (**D**) Western blot of DHX34 protein in WT cells compared with the heterozygous mutant from A. (**E**) qPCR of the expression of NMD target genes in WT cells compared with the heterozygous mutant from A. The average, SD, and P value (two-tailed t test) are from three independent experiments.

Splicing variations in NMD factor genes. The loss of functional RNA decay activity by increased inclusion of *DHX34* poison exon 12b is suggestive of the need for downregulation of NMD to sustain AML disease pathology. Therefore, I hypothesized that alternative splicing of NMD factors is also dysregulated among AML patients. To address this question, I repeated once more a MAJIQ analysis to query for highly variable splicing modules in curated list of well-documented NMD factors (**Figure 3.5A-C**). The most variable splicing module within NMD factor genes represents the alternative first exon usage of *SMG1* transcripts. Instead of truncating the protein as is the case with the insertion of PTC, alternative first exon usage could potentially change the function of the encoded protein. Protein kinase SMG1, a phosphoinositide 3-kinase (PI3K)-like kinase, phosphorylates UPF1 at multiple serine-threonine-glutamine ([S/T]Q) motifs in its C-terminal domain thereby activating UPF1-mediated NMD [150]. The use of an alternative first exon in *SMG1* is coupled with the use of a different 5'UTR, which can affect transcript regulation and localization. Furthermore, the intron sequence between the first exon and the alternative exon ~26 kilobases in length, suggesting the potential use of an alternative promoter to initiate transcription at the alternative first exon. The second most variable splicing event represents the alternative last exon inclusion of *PYM1*, which encodes a protein that interacts with EJC components MAGOH and RBM8A. The alternative last exon usage leads to switch a shorter isoform that also includes an entirely different 3'UTR which may confer a point of regulation at the level of mRNA localization and degradation. Additionally, it is thought that the N-terminus of PYM1 binds to both MAGOH and RBM8A [151], thus the translation of a truncated peptide can potentially exert dominant negative effects by sequestering active EJC complexes.

Interestingly, I did not find any highly variable splicing modules within the gene *UPF1*, which encodes the most essential factor required to disassemble the structured messenger ribonucleoproteins actively undergoing degradation [152]. However, I do find a highly variable splicing event in the gene *UPF2* representing the inclusion of a cryptic poison exon between exons 9 and 10 – which I termed exon 9b. The inclusion of poison exon 9b is expected to functionally downregulate protein similar to the inclusion of poison exon 12b in *DHX34*. Interestingly, the inclusion of *UPF2* exon 9b is present at steady state levels (~35%) in CD34 normal donors with a variability higher than the threshold defined in Figure 2.2, alluding to a potential feedback loop that may be regulating the expression of the encoded protein. Furthermore, it appears that the majority of the PENN-AML patients actually have significantly *less* inclusion of *UPF2* poison exon 9b compared to CD34 normal cells. UPF2 is an adaptor protein that brings together UPF1 and UPF3 to elicit NMD and there is contradicting evidence that UPF2 is dispensable for the successful NMD. Thus, the finding that *UPF2* poison exon 9b is significantly less included in AML patients compared to normal donors is intriguing, as it alludes to the idea that at least a subset of patients may depend on increased of UPF2 protein expression for AML blasts cell viability.

Additionally, I find splicing variations in *UPF3A* and not *UPF3B*. Specifically, the splicing variation in *UPF3A* represents the skipping of exon 8 which results in a frameshift alteration and creates downstream PTCs and is expected to functionally downregulate the encoded protein. The UPF3 protein is encoded by two paralogs: UPF3A and UPF3B. UPF3A serves as a weak NMD factor and NMD repressor by sequestering UPF2 from the NMD machinery, while UPF3B is a NMD branch-specific factor that directly binds the EJC

and stimulates NMD [153]. Therefore, it appears that highly variable splicing events in AML patients are targeting NMD factors with auxiliary-type involvement in NMD activity, instead of targeting essential regulators such as UPF1. This suggests that the AML blast do not tolerate splicing alterations to essential genes, and that they may be adapting the balance of the NMD-splicing regulatory axis to favor cell survival.

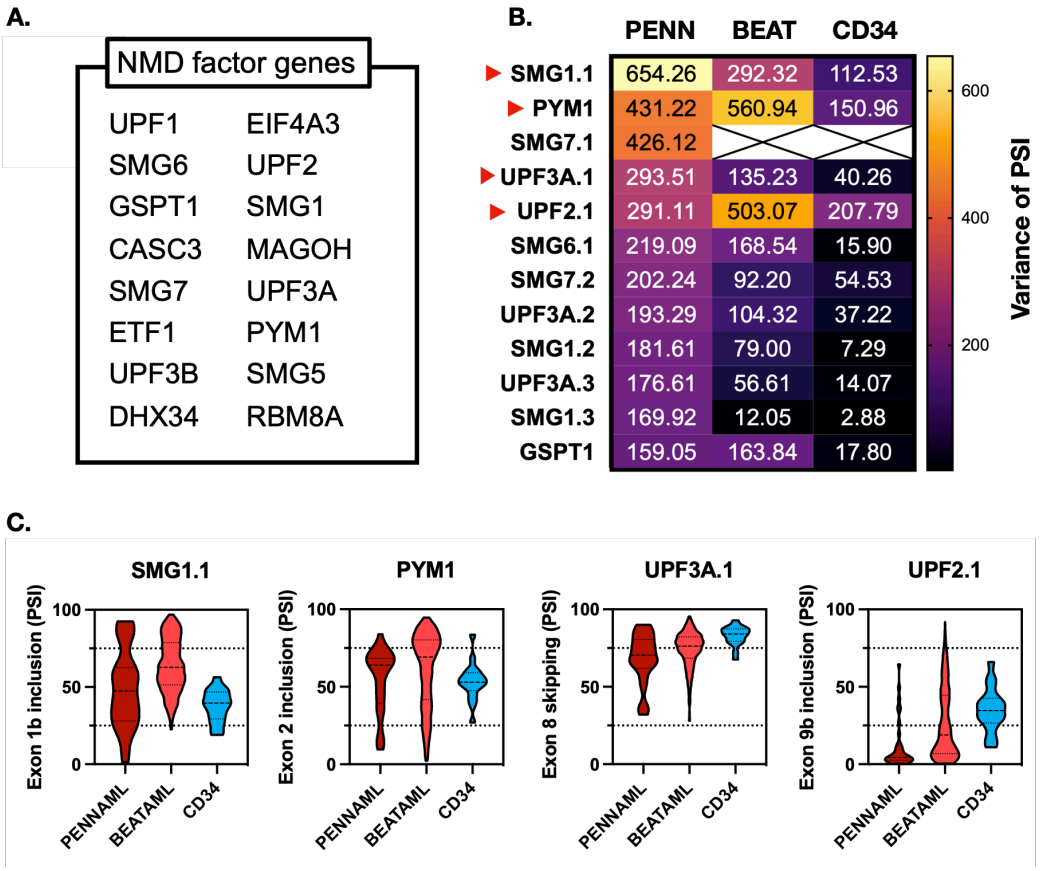


Figure 3.5 – **Highly variable splicing modules in NMD factor genes.** (**A**) List of NMD factor genes used in the extended analysis of splicing variations. (**B**) Heatmap showing the variance of highly variable splicing modules (Red arrows depict the splicing modules plotted in panel C) (Variance threshold cutoff is the same used in the initial analysis shown in Figure 2.2) (**C**) Violin plots showing PSI distribution of the most variable splicing modules in NMD factors across AML cohorts and CD34 normal donors.

Discussion

Collectively, the analyses discussed in Chapter 3 elucidate the pervasive role of alternative RNA splicing in regulating successful gene expression. Specifically, I discover a significant number of patients with higher inclusion of poison exon 12b in *DHX34* transcripts Furthermore, I demonstrate the function of this poison exon in regulating the successful expression of this gene through a previously undocumented feedback loop. I successfully inferred a subset of potential target substrates of DHX34 NMD activity using data from an independent study of NMD factors that had not directly tested for DHX34 targets. I experimentally confirmed using a CRISPR-engineered cell line model that the inferred target substrates of *DHX34* NMD activity were indeed altered upon forced inclusion of poison exon 12b. The DHX34 target gene substrates include transcripts encoding regulators of chromatin, such as *CHD2* and *METTL22*, and thus present a potential alternative avenue for further dysregulation of chromatin structure within AML blasts. Therefore, NMD inactivation that leads to the over expression of oncogenes and/or the perturbation of functional transcript expression of tumor suppressors may be an alternative route that supports AML disease pathology.

Interestingly, the inclusion of *DHX34* poison exon 12b seems to be present at steady state levels (~50%) in the CD34 donor cohort, suggesting its potential utility in regulating DHX34 protein expression as part of normal cell physiology. Of note, essential poison exons encoded within the genome tend to show higher conservation scores around the sequence boundaries of the particular exon occurring within the intronic sequence, however, that is not the case for cryptic exon DHX34 exon 12b, where there is absolutely no conservation between phylogenetic species. Questions still remain regarding the

particular regulators of poison exon 12b inclusion as well as the NMD factors that may target *DHX34* transcripts themselves for decay. Motif analysis of the exon sequences of *DHX34* exon 12, exon 12b and exon 13 as well as flanking intronic sequences reveal several UGGYUG motifs peppered across exons 12 and exon 12b. Furthermore, I also observe a large AGRCCANCAA motif within the boundaries of each of the three exons as well as several instances of AAG-rich sequences, suggesting potential competitive regulation by a poly-purine binder.

To understand the pervasiveness of alterations to the NMD pathway within AML patients, I expanded my analysis to include known regulators of NMD. I find that a subset of NMD factors indeed harbor highly variable splicing events that are predicted to functionally disrupt NMD. Although there is evidence of the coupling of splicing variations and NMD, the highly variable splicing events within NMD factor transcripts that I describe in the AML patients are undocumented. The alternative splicing events in *SMG1* and *PYM1* directly affect proteins involved in UPF1's interaction with the EJC as well as its activation via kinase phosphorylation. Additionally, the significant reduction of *UPF2* poison exon 9b inclusion is predicted to increase the availability of UPF2 protein in the AML blast cells. Similarly, the reduction of *UPF3A* exon 8 skipping is also predicted to increase the availability of UPF3A protein, a known repressor of UPF2. Thus, the splicing events in *UPF2* and *UPF3A* suggest that AML blast cells may rely UPF2 activity to sustain their altered cellular program. Collectively, my findings are consistent with the widely accepted coupling of alternative splicing and NMD activity in the cell and pose new questions as to how the cancer cell may be leveraging NMD dysregulation via the plasticity of alternative mRNA splicing to promote oncogenicity and cell viability.

CHAPTER IV

Conclusion, Perspectives and Future Directions

Conclusions

The work presented throughout this dissertation underscores the contribution of alternative RNA splicing to the functional dysregulation of particular genes that typically harbor mutations associated with the development and progression of AML. As discussed in Chapter 1, AML is an aggressive type of hematologic cancer characterize by the clonal proliferation of poorly differentiated myeloid progenitor cells. AML blasts usually harbor chromosomal rearrangements, cytogenetic abnormalities and gene fusions that confer aggressive oncogenic capacity. However, it is now understood that about half of AML patients have a cytogenetically normal karyotype and that genetic mutations such as single nucleotide substitutions and small tandem duplications account for the majority of driver molecular lesions in AML. Regardless of the significant progress in the study of AML, the genetic basis of this disease remains to be fully understood given that mutations in individual AML-associated genes are lowly penetrant across the patient population. In particular, most AML-associated genes are mutated in less than 10% of patients, with only three genes (*FLT3*, *NPM1*, *DNMT3A*) harboring mutations in more than 20% of the AML patient population. Furthermore, the patient population is significantly heterogeneous because most patients harbor mutations in more than 1 particular gene, and these mutations are expressed in complex patterns of co-occurrence and mutual exclusivity.

Importantly, as I have thoroughly discussed throughout this dissertation, genetic mutations are not the only way to dysregulate a gene. Specifically, my study highlights that alternative pre-mRNA splicing is a gene regulation mechanism capable of disrupting protein function that is largely underappreciated in the study of AML and of cancer. Of note, mutations in splicing factor genes are common in myeloid malignancies overall and

have been shown to induce different disease phenotypes according to cellular context. While a handful splicing factors (U2AF1, SRSF2, SF3B1, and ZRSR2) are among the so-called "commonly mutated" genes in AML, these also are present in less than 10% of patients and no common pattern of splicing dysregulation has been associated with these mutations. Therefore, the study that I have performed for this dissertation aimed to uncover variations in pre-mRNA splicing that were not only common across the AML population but also independent of any known splice factor mutation.

Alternative mRNA splicing outcomes can lead to the inclusion of exonic sequences that result in either the insertion of an in-frame PTC or the production of a *de novo* PTC via frame-shift alterations. However, although the potentially deleterious effects of dysregulated mRNA splicing are acknowledged in the literature, no particular study had focused on identifying splicing variations strictly within genes known to be commonly altered by genetic mutation in AML. To address this gap, I leveraged the use of poly(A)-selected mRNA-sequencing data to perform a MAJIQ high-throughput analysis of alternative splicing for a set of 70 AML-associated genes. I built an extended pipeline to processes MAJIQ splicing quantification data into distinct *splicing modules*, which is a data feature that I developed to facilitate filtering and dimensionality reduction in the analyses presented within this dissertation. Specifically, MAJIQ quantifies and models splicing events in terms of local splicing variations (LSVs), and these LSVs can sometimes overlap as well as point to the same molecular event. This merited the development of a new feature, namely *splicing modules*, that allowed me to combine overlapping LSVs and junctions into a distinct splicing event and produce relevant statistics. For each splicing module, the junction with the highest variance in percent spliced in (PSI) represented the intrinsic variability of the splicing choices of a particular event across AML patients.

To increase the likelihood of identifying functionally relevant splicing events, as well as to underscore the molecular heterogeneity of AML, I focused my analysis of on those splicing modules with the highest variability across the PENN-AML patients themselves. After identifying the most highly variable splicing modules present in the PENN-AML, I compared the intensity of the splicing module variances with that of a cohort of CD34+ normal myeloid donor cells as well as the large and independent BeatAML patient cohort (See Chapter 2, Figure 2.3). The quantification of splicing modules in $CD34^{+}$ donors revealed that these cells normally express a tight range of splicing decisions within this particular cell type (**Figure 4.1**). This suggests that healthy cells may be normally regulating these splicing events as part of intrinsic cell functions. Importantly, in most of the 40 highly variable splicing modules, the absolute change between the mean PSI for a module *m* in PENN-AML vs $CD34^{+}$ was less than 20%. This highlights the underlying heterogeneity of AML because a large proportion of patients express significantly different PSI in the splicing of events that on average may appear to be relatively unchanged.

Strikingly, the analyses performed in Chapter 2 uncovered that most of the highly variable splicing modules of interest are also highly correlated between themselves, which is suggestive of biological co-regulation (**Figure 4.1**). Specifically, I explored the relationship between splicing events, by producing a pairwise matrix of Spearman rank correlations between splicing modules and reordered the features by hierarchical clustering analysis which facilitated the identification of those splicing modules that exhibited the highest degree of co-regulation. This finding is of particular importance because it suggests that the penetrance of one particular splicing variation affecting an AML-associated gene can be coupled with the perturbation of multiple other AML-

associated genes via altered splicing. Furthermore, the co-regulation of the splicing modules found in PENN-AML was equally observed within the BeatAML patient cohort, suggesting that this pattern is representative of underlying biological mechanisms of AML pathology at large.

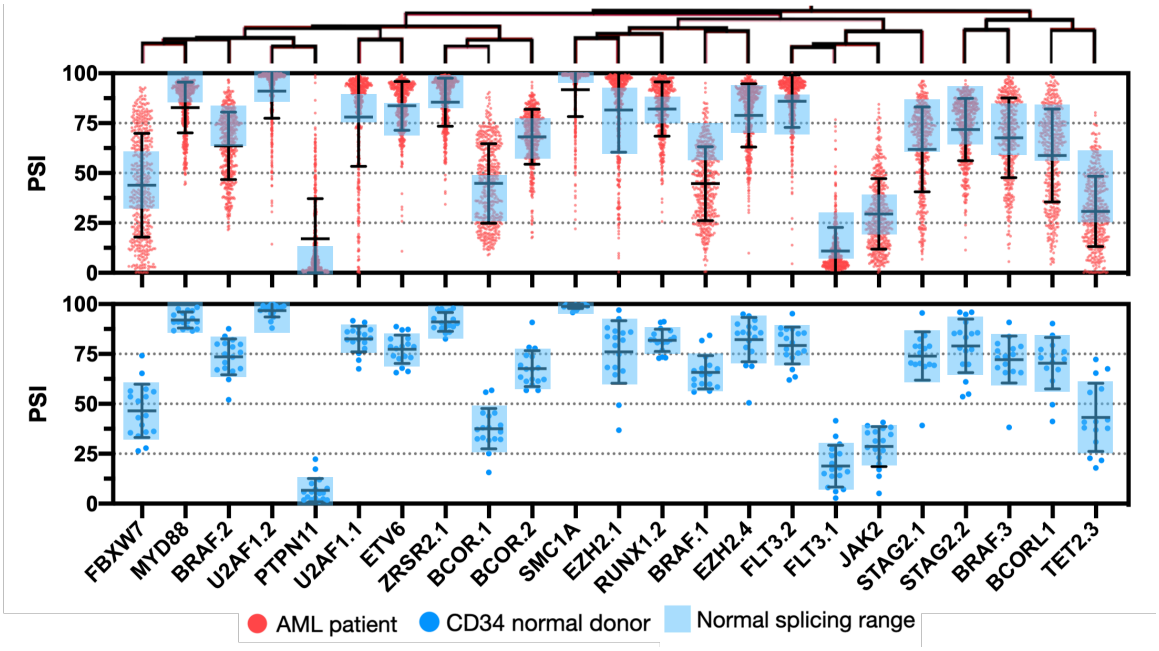


Figure 4.1 –**PSI distributions of the 23 highly co-regulated splicing modules across AML patients and CD34 normal donors**. The top panel is a violin plot of showing the pooled distribution of PSI for the coregulated splicing modules across PENN-AML and BeatAML patients. The bottom panel is PSI distribution in CD34 normal donor cells of the same modules in top panel. The violin plots are ordered based on the hierarchical clustering dendrogram displayed on the top panel. The line whiskers within the distribution plots represent the mean PSI ± 1 standard deviation for each splicing module in each panel. The normal range in splicing of $CD34^{+}$ cells was plotted in both panels to show how many AML patients are significantly deviating from the levels observed in these healthy cells.

Critically, in the present study I experimentally characterized via radiolabeled RT-PCR and Western blot analyses a set of highly variable and co-regulated splicing events that result in the reduction of full-length protein expression facilitated by the inclusion of

a PTC. In the case of the gene *EZH2*, I discover that the skipping of exons 11 and 12 results in a frameshift alteration that leads to the production of a PTC in the upstream exon. This particular splicing event involves more than two junctions, and thus is categorized as a complex splicing event that can only be quantified by computational frameworks, such as MAJIQ, that leverage the modeling of LSVs. Additionally, I find within *EZH2* transcripts the inclusion of exon 9b, which introduces an in-frame PTC and has been previously described in the literature to induce NMD via UPF1-sensitivity assays [92]. The inclusion of exon 9b seems to be present at steady state levels of around 25% within CD34 normal donors, and the high degree of sequence conservation within the intronic sequence suggests that this particular event is a functionally relevant molecular switch. Furthermore, the skipping of EZH2 exon 11 and 12 was not correlated with the inclusion of exon 9b, suggesting that there are distinct underlying mechanisms by which the splicing of these exon sequences is altered.

The identification of a highly correlated set of splicing variations raises questions regarding the mechanism by which these are manifested within the AML blasts. In my analysis, I identify splicing variations in transcripts encoding splice factors themselves, namely *ZRSR2* and *U2AF1*, which are highly correlated with the skipping of *EZH2* exon 11 and 12. Specifically, I found that an alternative last exon facilitates the transcription of a short isoform of ZRSR2 which consequently downregulates the expression of full-length protein – across the patients who are then also deficient for EZH2 protein. In contrast, U2AF1 protein expression are seen unaffected by the variation in exon 3 to exon 3' splicing that I uncover within AML blasts, presumably because although the event itself is highly variable, it is only functionally present in less than 25 % of transcripts. The functional downregulation of a splicing factor via an autoregulatory alternative splicing could be

regulating the splicing modules in question. To test this, I depleted ZRSR2 and U2AF1 factors in MV411 cells; however, this did not recapitulate the splicing patterns observed in of AML cells.

The co-regulated splicing events in question could be a result of a more complex combination of molecular alterations that are not recapitulated by only the depletion of U2AF1 and/or ZRSR2. In light of this, I analyzed the sequence regions around the highly variable splicing modules to find evidence of regulatory mechanisms, and I found an enrichment for AAG-repeats. Interestingly, I also found a very strong correlation between gene expression of *SRSF1*, a well-known RBP with a preference for poly-purines, and the PSI values of most co-regulated splicing modules. However, knockdown of *SRSF1* was also not sufficient to induce the co-regulated variations in splicing observed for *EZH2* and *ZRSR2*. Therefore, there is still the unanswered question regarding the mechanisms that may be collectively leading to the manifestation of the co-regulated variations in mRNA splicing within AML-associated genes that I have catalogued across AML patient cohorts.

To further explore the role of alternative splicing in dysregulating key genes within AML blasts, I describe in Chapter 3 an expanded analysis that includes a set of 20 genes recently published to have an association with germline cases of AML After using the MAJIQ pipeline to query splicing modules within this additional set of genes, I found of particular interest the inclusion of poison exon 12b within the gene *DHX34*. Prior work had failed to find enrichment of somatic mutations within DHX34 outside of two instances of familial AML and thus, this is the first study to inquire about reduced function of DHX34 protein through alternative splicing within AML blasts. Although inclusion of the *DHX34* poison exon doesn't strongly correlate with the main cluster of the other splicing modules, some of the patients that had higher skipping of *EZH2* exon 11 and 12 also had

noticeably higher inclusion of *DHX34* poison exon. Therefore, the reduced efficiency of NMD that results from the inclusion of *DHX34* poison exon may explain the increased accumulation of at least some of the splicing events that I describe in my study. To test the effects of inclusion of *DHX34* poison exon 12b I inferred potential NMD targets, and I experimentally validated using CRISPR gene editing that forced inclusion results in the increased abundance of target transcript substrates (*CHD2*, *PARP6*, *METTL22* and *WHAMM*). These genes have demonstrated roles in important cell functions such as chromatin regulation and cell cycle progression; thus, the dysregulation of the respective gene transcripts may promote oncogenicity of AML blasts. Therefore, this analysis represents the first study of regulation of DHX34 protein expression via inclusion of poison exon 12b and for the first time identifies that this exon is included in significantly higher proportions across the population of sporadic AML patients.

RNA-seq data is rich information that can be leveraged to decode multiple layers of biological regulation. The aligned short reads can be further processed to determine small nucleotide variations within RNA transcripts. An analysis of genetic mutation frequencies in the PENN-AML cohort using the RNA-seq method confirmed mutations at known target loci of interest (**Figure 4.2**). The frequency of mutations identified for the PENN-AML cohort slightly differs from that of the broad AML population due to an overrepresentation of *FLT3*-ITDs and *FLT3* mutations as well as an underrepresentation of mutations in genes such as *DNMT3A*. Due to intrinsic methodological limitations, more work needs to be done to accurately derive these genetic variants from the RNA-seq data, however, the data is sufficient to provide an idea of the level of dysregulation that may be attributed to genetic mutation alone. Importantly, using the same RNA-seq data we can account that for most of the AML-associated genes, changes in alternative splicing

noticeably increases the proportions of AML patients with molecular alterations (**Figure 4.2**). Interestingly, although there are mutations in splicing factors SRSF2 and SF3B1, MAJIQ did not quantify any variability at the level of splicing for these genes. Furthermore, analysis of *DHX34* sequence results in the quantification of uncatalogued genetic mutations that merit further investigation.

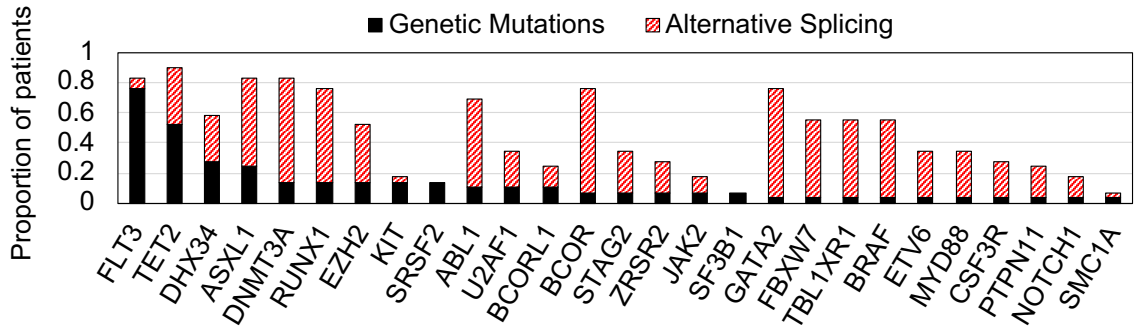


Figure 4.2 – **Proportion of PENN-AML patients with molecular alterations within AML-associated genes.** The black bar plot shows proportion of patients with genetic mutations quantified from the RNA-seq reads at particular loci of interest. Specifically, the proportion of FLT3 mutations also includes the proportion of FLT3-ITD which were determined via standard genotyping analysis. The red bar plot shows the proportion of additional patients that have molecular alterations in this case resulting from alternative RNA splicing. The proportion of patients with altered splicing was derived from the number of patients that expressed significantly different levels of a particular splicing module when compared to the normal range in splicing found in $CD34^{+}$ donor cells. The bar plot is sorted by decreasing frequency of genetic mutation followed by decreasing frequency of alternative splicing.

The identification of *DHX34* mis-expression in AML due to splicing not only underscores the importance of assessing splicing as a contributor to gene dysregulation in AML, but also highlights the potential importance of regulation of NMD to the pathology of this cancer. Therefore, building off of this particular finding and reasoning, I queried an additional set of splicing variations within NMD factor genes, in an effort to further understand the penetrance of functionally dysregulated NMD across the AML population.

I found significant variation in splicing of transcripts encoding core components of the decay inducing complex such as SMG1, UPF2 and UPF3A. Interestingly, I did not find variable splicing within transcripts encoding the most essential NMD factor, UPF1, presumably because alterations to the expression of this particular protein may not be tolerated by the AML blasts. Furthermore, this analysis also uncovered that some of the splicing events within NMD factor genes are strongly correlated with the splicing of particular AML-associated genes. Overall, my analyses and my findings further couple the regulation of NMD to alternative mRNA splicing and provides a deeper insight of the contribution of NMD to the pathology of AML and of cancer.

Perspectives & Future Directions

Acute myeloid leukemia is an aggressive and highly heterogenous form of hematologic cancer. There have been extensive studies regarding the genetics of AML, however, for at least half of the patient population, the molecular basis of the disease remains to be fully understood. As discussed in Chapter 1, there are a number of distinct genomic lesions that have been found within AML patients, leading to substantial disease heterogeneity. The central question motivating the work performed in this dissertation is whether mRNA splicing variations are an unaccounted source of molecular heterogeneity capable of leading to altered protein function. The mechanisms and outcomes of alternative splicing of individual transcripts are relatively well understood and the body of work presented throughout this dissertation further contributes to the knowledge base of alternative mRNA splicing as an essential contributor of functional gene expression. Through my query of splicing variations, I was able to identify that many more AML patients harbor molecular alterations that functionally dysregulate particular essential

proteins than what was normally being accounted for by just looking at activating and deleterious genetic mutations. Thus, my findings strongly underscore that alternative mRNA splicing needs to be quantified in the process of fully profiling the molecular basis of individual patients suffering from AML. Furthermore, I provide through my work a robust catalogue of alternative pre-mRNA splicing variations and I elucidate at subset of the most interesting events. However, more work needs to be done to fully understand the dysregulation of mRNA splicing in AML.

Critically, the depth and robustness of the high throughput splicing quantification analysis was facilitated by the MAJIQ computational framework and its use of the LSV concept. Other splicing quantification algorithms have intrinsic limitations that would have prevented me from properly identifying the more complex sets of splicing variations that I described in my study. As previously discussed, many of the high throughput splicing quantification algorithms methods that were available at the time of initiating my studies only analyzed known and annotated transcript isoforms. In contrast, MAJIQ supplements known transcripts with *de novo* junctions inferred from the RNA-seq data. Although I focused on elucidating the significance of *DHX34* exon 12b inclusion, which is a simple binary splicing event, most of the splicing modules that I describe include the use of unannotated cryptic exons as well as choice between more than 2 splicing choices. Therefore, the use of LSVs as well as the detection of *de novo* splice junctions by the MAJIQ framework is currently the most ideal tool in the field to generate a robust transcriptomic profile that accurately captures the complexity of the mRNA splicing choices elicited by cancer cells.

Most splicing quantification algorithms rely on a differential model that requires first the definition of a normal control. In contrast, the MAJIQ-PSI algorithm calculates

empirical distributions of junction usage across every single RNA-seq sample individually, and therefore facilitates the N-of-1 approach that is needed to study a highly heterogenous disease such as AML. I emphasize that I first identified splicing modules across each of the PENN-AML patients individually, and afterwards I compared the behavior of these splicing modules across the larger BeatAML cohort as well as the cohort of $CD34^+$ healthy donor cells. Thus, the MAJIQ splicing quantification algorithm produced a catalogue of mRNA splicing events, that includes those events with usage of junctions that is highly *variable* across PENN-AML patients, as well as those events with junction inclusion that is highly *constitutive* across patients. The robustness of this approach allowed me to go one step further in my analyses, and thus quantify and experimentally validate rare splicing events present only in 1 patient that are also expected to have a deleterious effect, which was the case of the skipping of exon 9 in *ZRSR2* (See Figure 2.6). These particular rare splicing events are ignored if I were to perform differential comparison of grouped AML patients vs. $CD34^+$ normal donor cells, which is the case of most high-throughput studies of splicing in AML.

Alternative splicing is naturally coordinated by RBPs that are capable of binding to multiple distinct mRNA transcripts and thus facilitate altered splicing of multiple target genes. Not surprisingly, recent efforts by other research groups have been directed more towards studying the coordination of alternative splicing networks by leveraging the use of short-read mRNA-sequencing data. Concordant with the field's efforts, the particular analysis that I have performed in this dissertation aims to provide a pipeline for the identification of co-regulated splicing events supported by the validation of such events in AML blasts. Specifically, the bioinformatics pipeline designed for this study predicted the co-occurrence of splicing events in *EZH2* and *ZRSR2* within the patient-derived AML

blasts, which we successfully validated via radiolabeled RT-PCR and western blot analysis. While the signature of co-regulated splicing events is particularly striking, more experiments should be performed to elucidate the mechanism by which this is manifested across AML patients. In particular, it proved to be rather difficult to recapitulate these splicing events in cell lines of leukemic origin. Transient knockdown of splice factors is difficult because the cell typically upregulates the expression of splicing factors to compensate for the lack of functional protein. Thus, an alternative approach could be to force the use of *ZRSR2* exon 3b and *SRSF1* poison exon 4 using CRISPR, and thus force the expression of the truncated isoforms and test for downstream changes in gene expression and alternative splicing.

The occurrence of multiple co-occurring splicing variations poses the question as to how the cell can adapt to support such a perturbed molecular profile, and if these splicing events are associated with a particular outcome. One particular splice factor can regulate hundreds of distinct splicing events, each of which may be contributing at varying degrees to the oncogenicity of AML blasts. The widespread effects of dysregulated splicing somewhat mirror the molecular two-hit model that has been historically proposed in AML, where a particular myeloid cell needs to have a block in differentiation as well as an activation of proliferation to elicit clonal expansion [154,155]. Specifically, the original concept of the "AML two-hit model of pathogenesis" was limited to the alterations of genes via genetic mutations and not via alternative mRNA splicing. As I have shown through my work, particular AML patients have a significant functional downregulation of multiple key proteins via dysregulated alternative mRNA splicing. I emphasize that the results of my studies contribute to the molecular profiling of an heterogenous AML population, as well as to the elucidation of mechanisms that may be potentially driving the oncogenicity

and pathology of AML blasts. Future studies should focus on understanding the molecular burden of the splicing events that I have identified throughout my studies.

Both *EZH2* and *ZRSR2,* as well as other AML-associated genes, harbored multiple distinct splicing variations that were un-correlated, which underscores the complexity of multiple splicing decisions within the same transcripts that can affect the expression of the respective gene. Furthermore, I note that uncorrelated variable splicing modules within the same gene transcripts appear to result in the same loss- or change-of-function effect, but through distinct avenues. The functional loss of EZH2 via skipping of exons 11 and 12 is expected to reduce cellular H3K27me di- or tri-methylation levels which consequently would de-repress the expression of EZH2 target genes. I specifically did not probe for downregulation of H3K27me di- or tri-methylation levels because of limited sample availability of the patient AML blasts. However, there is functional evidence that inclusion of *EZH2* poison exon 9b, which is uncorrelated with the skipping of exon 11 and 12, leads to the a hypomethylated state [92]. Thus, future analyses should include developing multivariate models that capture the distinct avenues that functionally downregulate a particular AML-associated gene. Specifically, an integrative model of molecular lesions that functionally downregulate EZH2 protein expression and activity should at least include genetic mutations such as D185H, alternative skipping of exon 11/12 and alternative inclusion of exon 9b. It remains a question as to which particular method is best suited to integrate fundamentally different data sources that represent distinct layers of biological regulation. However, ideally a multivariate regression model representative of EZH2 functional downregulation should be at least be predictive of dysregulation of known EZH2 target genes.

This body of work is also the first to describe the inclusion of *DHX34* poison exon 12b as a gene expression regulatory mechanism. The significant increase of inclusion within AML patient-derived blasts is suggestive of the need for the cancer cell to functionally downregulate this particular gene to sustain oncogenicity. Furthermore, the inclusion of this exon in 50% of DHX34 transcripts expressed in CD34+ donor cells suggest that this exon may be included as part of intrinsic regulatory mechanisms. As discussed thoroughly in Chapter 3, the DHX34 protein is involved in promoting NMD, and functional loss of this gene causes a significant increase in the abundance of target transcript substrates. The increased number of transcripts within the AML blasts is expected to promote cellular stress and interfere with RNA-RNA and RNA-protein interactions, which are essential for successful gene expression and regulation.

The role of NMD in carcinogenesis remains somewhat contradictory and unclear. Most studies of cancer focus on the role of NMD in regulating PTC-inducing mutations in oncogenes and tumor suppressor genes, rather than looking at downregulation of the NMD pathway per se. Not surprisingly, tumor-suppressor genes (for example *BRCA1*, *TP53, WT1*) are characterized by harboring more NMD-inducing nonsense mutations relative to oncogenes, which often contain missense mutations [156]. This highlights the crucial role of NMD in preventing the cell from translating truncated tumor suppressor transcripts and thus protecting cells from malignant growth. Interestingly, inactivating mutations in UPF1, UPF2, SMG1 and SMG6 are lethal, underscoring the essential role of NMD in normal cell development [157,158]. Specifically, UPF1 NMD activity has been shown to degrade particular pro-apoptotic factors so that cells avoid undergoing apoptosis [159]. Furthermore, during regulated apoptosis, caspases cleave the UPF1 and UPF2 proteins to shutdown NMD, highlighting a regulatory feedback loop between the suppression and

promotion of apoptosis [160]. This mechanism is conserved in human cells and likely helps to explain why NMD has been found to be essential for embryonic viability. However, this contradicts the observed inactivation of NMD in some tumors, suggesting that cancer cells may adapt to circumvent the consequent upregulation of pro-apoptotic signals that are targeted by NMD.

The loss of NMD pathway activity is expected to increase the burden of aberrant transcripts in the cell. NMD factors interact with the EJC to define 'prematureness' of stop codons, and normal transcripts escape NMD because all of the EJCs are typically deposited upstream of the stop codon, permitting the ribosome to displace them. In contrast, an aberrant transcript harboring a stop codon in a premature position will typically have at least one EJC downstream of the PTC, thereby triggering NMD. To keep the levels of physiological NMD targets constant, the efficiency of NMD itself is tightly regulated. Consistently, there is evidence of NMD factors being targeted themselves by the NMD pathway as a result of regulated alternative splicing. The persistence of aberrant transcripts within the cancer cell due to NMD inactivation may result in *de novo* RNA-RNA and RNA-RBP interactions that are likely capable of promoting widespread changes in gene expression regulation and RBP activity.

Specifically, non-coding RNAs are key posttranscriptional regulators of diverse cellular functions and cell fate determination, including hematopoiesis. Over the past several years, studies have recognized the potential of long non-coding RNAs (lncRNAs) and short micro RNAs (miRNAs) acting as either tumor suppressors or oncogenes depending on the targeted gene regulatory context [161;162]. As discussed in Chapter 1, distinctive miRNA profiles have been associated with subtypes of AML defined by cytogenic profiling [64], and loss of distinct microRNAs have been shown to lead to

leukemia in mouse models [163]. Therefore, the increased accumulation of transcripts due to NMD dysregulation in AML blast cells is presumably capable of sequestering important microRNAs and potentially lead to the gene dysregulation leukemogenesis. More work needs to be done to determine the molecular burden of NMD inactivation in the network activity of non-coding RNAs. However, mapping non-coding RNAs to predicted target sequences within AML-associated gene transcripts disrupted by alternative splicing may provide an idea of the effects of increased competitive endogenous RNAs across the AML patient population.

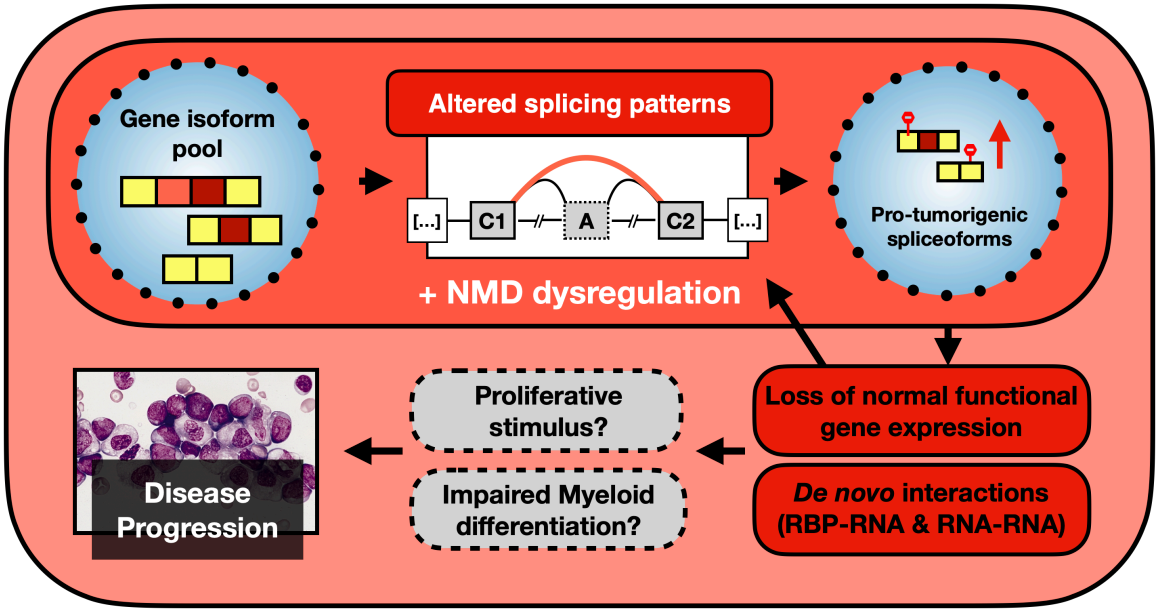


Figure 4.3 –**Model of oncogenic stimulus promoted by dysregulated mRNA splicing**. A an heterogenous pool of gene isoform transcripts is present within a cell at any given time. Altered splicing patterns may arise from changes to signal transduction as well as from mutations in splicing factors. The altered splicing patters increase the synthesis of non-productive transcripts as well as differential spliceoforms that translate distinct peptides. A downregulation of NMD presumably allows the persistent occurrence of non-functional transcripts. Collectively, the synthesis of these altered transcripts leads to a loss of functional gene expression that feeds back into the regulation of mRNA splicing to augment the pro-tumorigenic signal. Altered splicing patterns that result in loss- and change-of function effects presumably increase the proliferative capacity of cells while maintaining the differentiation blockade characteristic of AMLs, resulting in aggressive disease progression.

On a different note, many gene mutations that have been correlated to AML do not cause loss-of-function. For example, AML-linked genetic mutations in FLT3 and BRAF cause gain-of-function, while mutations in U2AF1 and ASXL1 alter protein function in ways that remain to be fully understood. While I have focused this study on splicing variations that introduce premature stop codons and thus reduce protein expression and function, several of the highly variable splicing modules I detect in the AML patients cause internal in-frame changes to the open reading frame or alter the N- and/or C-terminus of the encoded protein (**Table 2.1**). Of note, for over half of the genes that harbor gain- or change-of function genetic mutations in AML as well as highly variable splicing modules, the splicing is predicted to change the function of the protein rather than introduce a premature termination codon. This correlation suggests that these in-frame splicing variations may also phenocopy genetic mutations, as I have observed for the loss-of-function mutations. However, more work is required to fully elaborate the functional consequences of these in-frame splicing variations and compare their effect to genetic mutations in the same genes. An immediate experiment would involve again generating CRISPR models to test the effect of forced splicing in downstream functional gene activity similar to the experiments performed for *DHX34* poison exon 12b inclusion.

Closing Remarks

Through my work I aim to motivate larger studies of AML with clinically annotated datasets to determine if splicing patterns correlate with chemotherapeutic response and/or overall survival. I emphasize that my findings highlight the critical importance of assessing splicing patterns, in addition to genetic mutation and altered expression, in determining the contribution of individual genes to disease phenotype. While I demonstrate this to be true here for AML, I predict that assessment of splicing will be similarly relevant to the study of gene dysregulation in most other diseases.

Despite its heterogeneity, one common feature of AML is its capacity to evade programmed cell death and resist cytotoxic stimuli. Thus, 50% of elderly AML patients will resist treatment and die within the first 6 months of diagnosis, and over 40% of young patients relapse within 12 months of intensive chemotherapy [164,165]. The development of therapies that target resistance factors and/or promote cell death has therefore been a major focus of researchers and clinicians alike. The identification of splicing events that are broadly mis-regulated across the AML population provides important candidates to motivate further research aimed at testing the potential therapeutic relevance of manipulating these AML-associated splicing events. In in at least one instance of cancer preclinical models, pharmacological manipulation of mis-regulated exon inclusion splicing events via antisense oligonucleotide treatment has been used to decrease tumor growth [166].

In sum, short-read sequencing data originating from the RNA-seq methods used in this study is rich in information that can be leveraged to decode multiple layers of biological regulation. In this dissertation, I focused on surveying splicing variations across AML-associated genes using RNA-seq data from AML patients. I experimentally validated

splicing events in multiple distinct genes, and I uncovered new avenues by which AML-associated genes get functionally downregulated in AML. Importantly. I began elucidating how *DHX34* gets dysregulated via poison exon inclusion, a mechanism that has not been previously described in the literature. This finding then motivated the query of splicing variations in NMD factor genes, a direct study that had not been performed, which uncovered more interesting splicing events, and further coupled the activity of NMD with the fidelity of splicing. To conclude, I emphasize the importance of quantifying alternative splicing variations to paint a more accurate picture of the molecular pathology that drives the development not only of AML, but also of other cancer types and of human diseases at large.

METHODS

AML Patient Samples

To construct the PENN dataset, AML blasts were first isolated from AML patients via size-exclusion centrifugation apheresis or from peripheral blood on Ficoll gradients and stored in the Penn Stem Cell and Xenograft Core at the University of Pennsylvania. To promote the molecular characterization of a homogenous cell population, AML patient samples were confirmed to be at least 90% blasts. CD34+ cells from PENN were isolated from adult bone marrow by the Penn High Throughput Screening Core directed by Dr. Sara Cherry and Dr. David Schultz. Total RNA was isolated from AML blasts by Trizol followed by DNAse treatment and reprecipitation. Poly(A) selection, library preparation, and paired-end sequencing was performed by GeneWiz. Sample biospecimen data table was published as part of Rivera et al 2021, Table S1) [38]

Poly-A mRNA Sequencing and Data Processing

High-throughput mRNA Sequencing files (.fastq) were pre-processed by Trim Galore (https://github.com/FelixKrueger/TrimGalore) to trim short-read adapter sequences from sample reads. Significantly larger datasets, such as the Beat-AML cohort were pre-processed with BBDuk (Decontamination Using Kmers, https://jgi.doe.gov/data-and-tools/bbtools/). Pre-processed reads were then aligned to the human genome (GRCh38 Genome Reference Consortium Human Reference 38, GCA_000001405.15) using STAR version 2.5.4b (https://github.com/alexdobin/STAR). The read-depth of RNA-Seq data from the Penn cohort ranged from 50 to 200 million reads per sample, with read lengths ranging from 75-150 nucleotides.

mRNA Splicing Quantification, Alternative Splicing Detection and Visualization

RNA-seq reads mapping to 70 AML-related genes were used for downstream splicing quantification analyses. The MAJIQ framework (https://majiq.biociphers.org/) [17] was used to identify all local splicing variations (LSVs) within the defined sets of genes (e.g. AML-related genes and NMD factor genes) across the PENN-AML cohort. An LSV is defined as a single exon alternatively joined to more than one RNA segment. Junctions involved in an LSV are quantified by Percent Splicing Index (PSI) values that represent the relative fraction of poly-A selected mRNA transcripts in which the aforementioned exon is joined to each of the alternative RNA segments. Variations in junction PSI represent the occurrence of a particular splicing pattern across RNA-seq data. To summarize our data, the quantified LSVs were used to construct splicing modules which model the splicing events that are occurring within AML-related gene transcripts. For every splicing module, the variance in PSI for junction j, or var(PSI_j), across a particular cohort was calculated and used to represent the variability of a splicing module across patients in said cohort. Splicing modules with a var(PSI_j) higher than one standard deviation away of the mean var(PSI_j) seen across patients in PENN-AML were deemed as "highly-variable" and characterized in downstream analyses. Splicegraphs representative of alternative splicing variation were visualized using VIOLA, an HTML5 web-tool. The code repository of the pipeline developed for this study currently on Bitbucket and is available upon request.

Validation of Patient Donor Cohorts

Splicing modules found in the PENN-HTSC cohort were also queried within data from an independent cohort of AML patients and a healthy myeloid donor cohort. A total of 444 RNA-seq files from AML patients were downloaded from the Beat-AML study [95]. Moreover, patient RNA-seq files from a cohort of $CD34^+$ cells isolated from 17 healthy myeloid donors were downloaded from Leucegene (https://leucegene.ca/). RNA-seq data was processed as described above. For Western blots and RT-PCR, new aliquots of cells from the same patient collection were thawed and viable blasts harvested for RNA (Trizol) and western blot (RIPA lysis buffer).

RNA-seq Single Nucleotide Variant Calling

Patient RNA-seq files were subjected to base quality recalibration, a data pre- processing step that detects systematic errors made by the sequencer when it estimates the quality score of each base call (https://software.broadinstitute.org/gatk/). Recalibrated RNA-Seq reads were visualized within the Broad Institute Integrative Genomic Viewer (IGV_2.6.3). Clinically relevant genomic regions for AML-genes were provided by the Penn Center for Personalized Diagnostics (CPD). Single nucleotide variations (SNVs) for each AML patient within clinically relevant regions were called from their respective RNA-seq reads. I also screened for other known SNVs that have been previously tied to AML pathology. SNVs were required to be evidenced a sequencing depth larger than 10 reads.

Gene Expression and Isoform Measurements

Transcript-level quantification of gene expression was estimated from the RNA-seq data using Salmon quasi-mapping algorithm (https://combine-lab.github.io/salmon/).

Effective counts library sizes were computed using the edgeR package to account for composition biases between samples. Transcript quantification was normalized to gene-level reads per kilobase per million (RPKMs) and transformed to a measure of relative gene expression (log_2RPKM) in downstream analyses. Variance of relative gene expression associated with batch effects was regressed out of our transcriptomic models using the ComBat framework (https://rdrr.io/bioc/sva/man/ComBat.html).

Statistical Analyses

Splicing modules:

We plotted the distribution of splicing module variances across the cohorts and assed statistically significant differences between the mean splicing module variances using Student's t-test. A pairwise analysis between the PSI values of each "highly-variable" splicing module generated a dissimilarity matrix of Spearman's rho correlations. To account for the unknown vector directionality of splicing modules, I used the absolute value of the Spearman's rank correlation coefficient in our pairwise dissimilarity matrix.

Complete linkage hierarchical clustering of the Spearman rho dissimilarity matrix was used to cluster highly correlated splicing modules into a distinct group. Splicing modules grouped together via the hierarchical clustering algorithm were used to shed light into molecular co-regulation of alternative splicing. Pearson R test was used to determine the correlation coefficients between splicing modules and gene expression vectors.

Correlation of DHX34 splicing with NMD target expression:

We used a common threshold of p-value ≤ 0.05 on the correlation between DHX34 exon 12b inclusion ratio and the transcript abundance of NMD target substrates on both

datasets. The threshold over the p-value rather than correlation coefficient was used to account for the large difference in sample sizes in the two datasets. The above test based on random permutation of the observed expression values allows us to detect significant correlations while accounting for sample sizes. However, it is still possible that DHX34 expression values will significantly correlate with many genes in general. In order to assess the enrichment of significance correlation between DHX34 and the previously reported NMD targets I correlated DHX34 exon 12b inclusion with the transcript abundance of a set of 1000 random genes, and I bootstrapped the correlation results 1000 times, each time picking a new set of random set of genes. For each gene set randomization, I computed the number of correlations passing the above threshold in both datasets, which in turn allowed us to assess the p-value of significantly positively and negatively correlated genes in our original set of NMD targets using a one sample t-test as shown in Fig. 5E.

Cell Lines and CRISPR mutations

In vitro growth of MV411 (ATCC CRL-9591) and HL60 (ATCC CCL-240) myeloid cell lines were by standard methods. Due to relative ease of transfection and basal level of DHX34 exon 12b inclusion, CRISPR mutations in DHX34 were done in the lymphoma Jurkat cell line by Dr. Rakesh Chatrikhi, and grown as described in [167]. CRISPR mutations were done by transfecting cells with SpCas9-2A-GFP (pX458 plasmid, Addgene: https://www.addgene.org/48138) containing templates for guide RNAs, together with the repair template shown in Figure 3.4A. 48 hours after transfection cells were washed and sorted for GFP+ cells. Single cells were cultured in 96 well plates for 3 weeks, expanded and screened by PCR to identify modified clones. Primers used for genomic PCR were the same as for RT-PCR shown in Table S2.

Western blots, qPCR and radiolabeled RT-PCR

Radio-labeled radiolabeled RT-PCR was performed as previously described (30). Splicing modules were amplified and validated experimentally using the primer pairs specified in Table S2. For qPCR, total RNA was converted to cDNA via reverse transcription using MMLV reverse transcriptase, followed by amplification using the Power SYBR Green PCR Master Mix (Applied Biosystems) in a LightCycler® 96 Instrument (Roche Life Sciences) at 55°C for 10 min, 95°C for 1 min, and then 40 cycles of 95°C for 10 sec and 60°C for 1 min. Primers targeting genes of interest were designed over splice junctions and spanning regions of constitutive splicing. Amplicons were designed to be less than 200 bp in length. Radiolabeled RT-PCR and qPCR primer sequences are published in Table S2 of Rivera et al 2021 [38]. Antibodies used for Western blots were as follows: EZH2 (a kind gift from Dr. Roberto Bonasio, UPenn), U2AF35 (Abcam catalog ab86305), ZRSR2 (Novus NBP1-57307), DHX34 (Novus NBP1-91832), Lamin B1(Abcam ab133741), and GAPDH (Abcam ab128915).

REFERENCES

1. Frith, M. C. *et al.* A code for transcription initiation in mammalian genomes. *Genome Res.* **18,** 1–12 (2008).
2. Millhouse, S. & Manley, J. L. The C-terminal domain of RNA polymerase II functions as a phosphorylation-dependent splicing activator in a heterologous protein. *Mol. Cell. Biol.* **25,** 533–544 (2005).
3. Matera, A. G. & Wang, Z. A day in the life of the spliceosome. *Nat Rev Mol Cell Biol* **15,** 108–121 (2014).
4. Turunen, J. J., Niemelä, E. H., Verma, B. & Frilander, M. J. The significant other: splicing by the minor spliceosome. *WIREs RNA* **4,** 61–76 (2013).
5. Wahl, M. C., Will, C. L. & LUhrmann, R. The spliceosome: design principles of a dynamic RNP machine. *Cell* **136,** 701–718 (2009).
6. Naftelberg, S., Schor, I. E., Ast, G. & Kornblihtt, A. R. Regulation of alternative splicing through coupling with transcription and chromatin structure. *Annu. Rev. Biochem.* **84,** 165–198 (2015).
7. Hertel, K. J. Combinatorial control of exon recognition. *J. Biol. Chem.* **283,** 1211–1215 (2008).
8. Matlin, A. J., Clark, F. & Smith, C. W. J. Understanding alternative splicing: towards a cellular code. *Nat Rev Mol Cell Biol* **6,** 386–398 (2005).
9. Pandit, S. *et al.* Genome-wide analysis reveals SR protein cooperation and competition in regulated splicing. *Molecular Cell* **50,** 223–235 (2013).
10. Fu, X.-D. & Ares, M. Context-dependent control of alternative splicing by RNA-binding proteins. *Nature Publishing Group* **15,** 689–701 (2014).
11. Ule, J. & Blencowe, B. J. Alternative Splicing Regulatory Networks: Functions, Mechanisms, and Evolution. *Molecular Cell* **76,** 329–345 (2019).
12. Barash, Y. *et al.* Deciphering the splicing code. *Nature* **465,** 53–59 (2010).
13. Zhang, J., Kuo, C. C. J. & Chen, L. GC content around splice sites affects splicing through pre-mRNA secondary structures. *BMC Genomics* **12,** 90 (2011).
14. Lovci, M. T. *et al.* Rbfox proteins regulate alternative mRNA splicing through evolutionarily conserved RNA bridges. *Nat. Struct. Mol. Biol.* **20,** 1434–1442 (2013).
15. Pan, Q., Shai, O., Lee, L. J., Frey, B. J. & Blencowe, B. J. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nat. Genet.* **40,** 1413–1415 (2008).
16. Baralle, F. E. & Giudice, J. Alternative splicing as a regulator of development and tissue identity. *Nat Rev Mol Cell Biol* **18,** 437–451 (2017).
17. Vaquero-Garcia, J. *et al.* A new view of transcriptome complexity and regulation through the lens of local splicing variations. *Elife* **5,** e11752 (2016).
18. Papasaikas, P., Tejedor, J. R., Vigevani, L. & Valcárcel, J. Functional splicing network reveals extensive regulatory potential of the core spliceosomal machinery. *Molecular Cell* **57,** 7–22 (2015).

19. Ray, D. *et al.* A compendium of RNA-binding motifs for decoding gene regulation. *Nature* **499,** 172–177 (2013).
20. Dominguez, D. *et al.* Sequence, Structure, and Context Preferences of Human RNA Binding Proteins. *Molecular Cell* **70,** 854–867.e9 (2018).
21. Anczuków, O. & Krainer, A. R. Splicing-factor alterations in cancers. *RNA* **22,** 1285–1301 (2016).
22. Cherry, S. & Lynch, K. W. Alternative splicing and cancer: insights, opportunities, and challenges from an expanding view of the transcriptome. *Genes Dev.* **34,** 1005–1016 (2020).
23. Fu, X.-D. The superfamily of arginine/serine-rich splicing factors. *RNA* 663–680 (1995).
24. Long, J. C. & Caceres, J. F. The SR protein family of splicing factors: master regulators of gene expression. *Biochem J* **417,** 15–27 (2009).
25. Zhou, Z. & Fu, X.-D. Regulation of splicing by SR proteins and SR protein-specific kinases. *Chromosoma* **122,** 191–207 (2013).
26. Piñol-Roma, S., Choi, Y. D., Matunis, M. J. & Dreyfuss, G. Immunopurification of heterogeneous nuclear ribonucleoprotein particles reveals an assortment of RNA-binding proteins. *Genes Dev.* **2,** 215–227 (1988).
27. Lunde, B. M., Moore, C. & Varani, G. RNA-binding proteins: modular design for efficient function. *Nat Rev Mol Cell Biol* **8,** 479–490 (2007).
28. Hentze, M. W., Castello, A., Schwarzl, T. & Preiss, T. A brave new world of RNA-binding proteins. *Nature Publishing Group* **19,** 327–341 (2018).
29. Climente-González, H., Porta-Pardo, E., Godzik, A. & Eyras, E. The Functional Impact of Alternative Splicing in Cancer. *CellReports* **20,** 2215–2226 (2017).
30. Kahles, A. *et al.* Comprehensive Analysis of Alternative Splicing Across Tumors from 8,705 Patients. *Cancer Cell* **34,** 1–38 (2018).
31. Supek, F., Miñana, B., Valcárcel, J., Gabaldón, T. & Lehner, B. Synonymous mutations frequently act as driver mutations in human cancers. *Cell* **156,** 1324–1335 (2014).
32. Jung, H. *et al.* Intron retention is a widespread mechanism of tumor-suppressor inactivation. *Nat. Genet.* **47,** 1242–1248 (2015).
33. Jayasinghe, R. G. *et al.* Systematic Analysis of Splice-Site-Creating Mutations in Cancer. *CellReports* **23,** 270–281.e3 (2018).
34. Smith, C. W., Patton, J. G. & Nadal-Ginard, B. Alternative splicing in the control of gene expression. *Annu Rev Genet* **23,** 527–577 (1989).
35. Das, S. & Krainer, A. R. Emerging functions of SRSF1, splicing factor and oncoprotein, in RNA metabolism and cancer. *Mol. Cancer Res.* **12,** 1195–1204 (2014).
36. Akerman, M. *et al.* Differential connectivity of splicing activators and repressors to the human spliceosome. *Genome Biol.* **16,** 119 (2015).
37. Anczuków, O. *et al.* The splicing factor SRSF1 regulates apoptosis and proliferation to promote mammary epithelial cell transformation. *Nat. Struct. Mol. Biol.* **19,** 220–228 (2012).

38. Rivera, O. D. *et al.* Alternative splicing redefines landscape of commonly mutated genes in acute myeloid leukemia. *Proc. Natl. Acad. Sci. U.S.A.* **118,** (2021).
39. Sweetser, D. A. *et al.* Delineation of the minimal commonly deleted segment and identification of candidate tumor-suppressor genes in del(9q) acute myeloid leukemia. *Genes Chromosomes Cancer* **44,** 279–291 (2005).
40. Gallardo, M. *et al.* Aberrant hnRNP K expression: All roads lead to cancer. *Cell Cycle* **15,** 1552–1557 (2016).
41. Sotillo, E. *et al.* Convergence of Acquired Mutations and Alternative Splicing of CD19 Enables Resistance to CART-19 Immunotherapy. *Cancer Discov* **5,** 1282–1295 (2015).
42. Stark, M., Wichman, C., Avivi, I. & Assaraf, Y. G. Aberrant splicing of folylpolyglutamate synthetase as a novel mechanism of antifolate resistance in leukemia. *Blood* **113,** 4362–4369 (2009).
43. Hicks, M. J., Lam, B. J. & Hertel, K. J. Analyzing mechanisms of alternative pre-mRNA splicing using in vitro splicing assays. *Methods* **37,** 306–313 (2005).
44. Amarasinghe, S. L. *et al.* Opportunities and challenges in long-read sequencing data analysis. *Genome Biol.* **21,** 30 (2020).
45. Tang, A. D. *et al.* Full-length transcript characterization of SF3B1 mutation in chronic lymphocytic leukemia reveals downregulation of retained introns. *Nat Commun* **11,** 1438 (2020).
46. Alamancos, G. P., Pagès, A., Trincado, J. L., Bellora, N. & Eyras, E. Leveraging transcript quantification for fast computation of alternative splicing profiles. *RNA* **21,** 1521–1531 (2015).
47. Shen, S. *et al.* rMATS: robust and flexible detection of differential alternative splicing from replicate RNA-Seq data. *Proc. Natl. Acad. Sci. U.S.A.* **111,** E5593–601 (2014).
48. Katz, Y., Wang, E. T., Airoldi, E. M. & Burge, C. B. Analysis and design of RNA sequencing experiments for identifying isoform regulation. *Nat Methods* **7,** 1009–1015 (2010).
49. Wu, J. *et al.* SpliceTrap: a method to quantify alternative splicing under single cellular conditions. *Bioinformatics* **27,** 3010–3016 (2011).
50. Shi, Y. & Jiang, H. rSeqDiff: detecting differential isoform expression from RNA-Seq data using hierarchical likelihood ratio test. *PLoS ONE* **8,** e79448 (2013).
51. Trapnell, C. *et al.* Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat Protoc* **7,** 562–578 (2012).
52. Singh, D. *et al.* FDM: a graph-based statistical method to detect differential transcription using RNA-seq data. *Bioinformatics* **27,** 2633–2640 (2011).
53. Hu, Y. *et al.* DiffSplice: the genome-wide detection of differential splicing events with RNA-seq. *Nucleic Acids Res.* **41,** e39 (2013).
54. Anders, S., Reyes, A. & Huber, W. Detecting differential usage of exons from RNA-seq data. *Genome Res.* **22,** 2008–2017 (2012).
55. Jagannathan-Bogdan, M. & Zon, L. I. Hematopoiesis. *Development* **140,** 2463–2467 (2013).

56. Gazit, R. *et al.* Transcriptome analysis identifies regulators of hematopoietic stem and progenitor cells. *Stem Cell Reports* **1,** 266–280 (2013).
57. Seita, J. & Weissman, I. L. Hematopoietic stem cell: self-renewal versus differentiation. *Wiley Interdiscip Rev Syst Biol Med* **2,** 640–653 (2010).
58. Warner, J. K., Wang, J. C. Y., Hope, K. J., Jin, L. & Dick, J. E. Concepts of human leukemic development. *Oncogene* **23,** 7164–7177 (2004).
59. Chen, J., Odenike, O. & Rowley, J. D. Leukaemogenesis: more than mutant genes. *Nat Rev Cancer* **10,** 23–36 (2010).
60. Döhner, H., Weisdorf, D. J. & Bloomfield, C. D. Acute Myeloid Leukemia. *N. Engl. J. Med.* **373,** 1136–1152 (2015).
61. Vardiman, J. W., Harris, N. L. & Brunning, R. D. The World Health Organization (WHO) classification of the myeloid neoplasms. *Blood* **100,** 2292–2302 (2002).
62. Bradstock, K., Matthews, J., Benson, E., Page, F. & Bishop, J. Prognostic value of immunophenotyping in acute myeloid leukemia. Australian Leukaemia Study Group. *Blood* **84,** 1220–1225 (1994).
63. Raspadori, D., Lauria, F., Ventura, M. A., research, D. R. L.1997. Incidence and prognostic relevance of CD34 expression in acute myeloblastic leukemia: analysis of 141 cases. *Elsevier*
64. Cancer Genome Atlas Research Network *et al.* Genomic and epigenomic landscapes of adult de novo acute myeloid leukemia. *N. Engl. J. Med.* **368,** 2059–2074 (2013).
65. Kuykendall, A., Duployez, N., Boissel, N., Lancet, J. E. & Welch, J. S. Acute Myeloid Leukemia: The Good, the Bad, and the Ugly. *Am Soc Clin Oncol Educ Book* **38,** 555–573 (2018).
66. Reilly, J. T. Pathogenesis of acute myeloid leukaemia and inv(16)(p13;q22): a paradigm for understanding leukaemogenesis? *Br. J. Haematol.* **128,** 18–34 (2005).
67. Grimwade, D. *et al.* The importance of diagnostic cytogenetics on outcome in AML: analysis of 1,612 patients entered into the MRC AML 10 trial. The Medical Research Council Adult and Children's Leukaemia Working Parties. *Blood* **92,** 2322–2333 (1998).
68. Chessells, J. M., Harrison, C. J., Kempski, H., Leukemia, D. W.2002. Clinical features, cytogenetics and outcome in acute lymphoblastic and myeloid leukaemia of infancy: report from the MRC Childhood Leukaemia working .... *nature.com* doi:10.1038/sj/leu/2402468
69. Papaemmanuil, E. *et al.* Genomic Classification and Prognosis in Acute Myeloid Leukemia. *N. Engl. J. Med.* **374,** 2209–2221 (2016).
70. Corces-Zimmerman, M. R., Hong, W.-J., Weissman, I. L., Medeiros, B. C. & Majeti, R. Preleukemic mutations in human acute myeloid leukemia affect epigenetic regulators and persist in remission. *Proc. Natl. Acad. Sci. U.S.A.* **111,** 2548–2553 (2014).
71. Shlush, L. I. *et al.* Identification of pre-leukaemic haematopoietic stem cells in acute leukaemia. *Nature* **506,** 328–333 (2014).

72. Yoshimi, A. *et al.* Coordinated alterations in RNA splicing and epigenetic regulation drive leukaemogenesis. *Nature* 1–34 (2019). doi:10.1038/s41586-019-1618-0
73. Grafone, T., Palmisano, M., Nicci, C. & Storti, S. An overview on the role of FLT3-tyrosine kinase receptor in acute myeloid leukemia: biology and treatment. *Oncol Rev* **6,** e8 (2012).
74. Small, D. *et al.* STK-1, the human homolog of Flk-2/Flt-3, is selectively expressed in CD34+ human bone marrow cells and is involved in the proliferation of early progenitor/stem cells. *Proc Natl Acad Sci USA* **91,** 459–463 (1994).
75. Griffith, J. *et al.* The structural basis for autoinhibition of FLT3 by the juxtamembrane domain. *Molecular Cell* **13,** 169–178 (2004).
76. Nguyen, B. *et al.* FLT3 activating mutations display differential sensitivity to multiple tyrosine kinase inhibitors. *Oncotarget* **8,** 10931–10944 (2017).
77. Chung, K. Y. *et al.* Enforced expression of an Flt3 internal tandem duplication in human CD34+ cells confers properties of self-renewal and enhanced erythropoiesis. *Blood* **105,** 77–84 (2005).
78. Levis, M. FLT3 mutations in acute myeloid leukemia: what is the best approach in 2013? *Hematology Am Soc Hematol Educ Program* **2013,** 220–226 (2013).
79. Hingorani, K., Szebeni, A. & Olson, M. O. Mapping the functional domains of nucleolar protein B23. *J. Biol. Chem.* **275,** 24451–24457 (2000).
80. Okuda, M. *et al.* Nucleophosmin/B23 is a target of CDK2/cyclin E in centrosome duplication. *Cell* **103,** 127–140 (2000).
81. Kurki, S. *et al.* Nucleolar protein NPM interacts with HDM2 and protects tumor suppressor protein p53 from HDM2-mediated degradation. *Elsevier* doi:10.1016/s1535-6108(04)00110-2
82. Chou, S.-H. *et al.* A knock-in Npm1 mutation in mice results in myeloproliferation and implies a perturbation in hematopoietic microenvironment. *PLoS ONE* **7,** e49769 (2012).
83. Sood, R., Kamikubo, Y. & Liu, P. Role of RUNX1 in hematological malignancies. *Blood* **129,** 2070–2082 (2017).
84. Osato, M. Point mutations in the RUNX1/AML1 gene: another actor in RUNX leukemia. *Oncogene* **23,** 4284–4296 (2004).
85. Preudhomme, C. *et al.* High incidence of biallelic point mutations in the Runt domain of the AML1/PEBP2 alpha B gene in Mo acute myeloid leukemia and in myeloid malignancies with acquired trisomy 21. *Blood* **96,** 2862–2869 (2000).
86. Ley, T. J. *et al.* DNMT3A mutations in acute myeloid leukemia. *N. Engl. J. Med.* **363,** 2424–2433 (2010).
87. Milne, T. A. *et al.* MLL targets SET domain methyltransferase activity to Hox gene promoters. *Molecular Cell* **10,** 1107–1117 (2002).
88. Margueron, R. & Reinberg, D. The Polycomb complex PRC2 and its mark in life. *Nature* **469,** 343–349 (2011).
89. Göllner, S. *et al.* Loss of the histone methyltransferase EZH2 induces resistance to multiple drugs in acute myeloid leukemia. *Nature Medicine* **23,** 69–78 (2017).

90. Hahn, C. N., Venugopal, P., Scott, H. S. & Hiwase, D. K. Splice factor mutations and alternative splicing as drivers of hematopoietic malignancy. *Immunol. Rev.* **263,** 257–278 (2015).
91. Papaemmanuil, E. *et al.* Clinical and biological implications of driver mutations in myelodysplastic syndromes. *Blood* **122,** 3616–27– quiz 3699 (2013).
92. Kim, E. *et al.* SRSF2 Mutations Contribute to Myelodysplasia by Mutant-Specific Effects on Exon Recognition. *Cancer Cell* **27,** 617–630 (2015).
93. Fei, D. L. *et al.* Impaired hematopoiesis and leukemia development in mice with a conditional knock-in allele of a mutant splicing factor gene U2af1. *Proc. Natl. Acad. Sci. U.S.A.* **115,** E10437–E10446 (2018).
94. Wang, H. *et al.* Prognostic value of U2AF1 mutant in patients with de novo myelodysplastic syndromes: a meta-analysis. *Ann Hematol* **98,** 2629–2639 (2019).
95. Tyner, J. W. *et al.* Functional genomic landscape of acute myeloid leukaemia. *Nature* **562,** 1–27 (2018).
96. Will, C. L. & LUhrmann, R. Spliceosome structure and function. *Cold Spring Harb Perspect Biol* **3,** a003707 (2011).
97. Gozani, O., Potashkin, J. & Reed, R. A potential role for U2AF-SAP 155 interactions in recruiting U2 snRNP to the branch site. *Mol. Cell. Biol.* **18,** 4752–4760 (1998).
98. Zhang, J. *et al.* Disease-Causing Mutations in SF3B1 Alter Splicing by Disrupting Interaction with SUGP1. *Molecular Cell* **76,** 82–95.e7 (2019).
99. Dvinge, H., Kim, E., Abdel-Wahab, O. & Bradley, R. K. RNA splicing factors as oncoproteins and tumour suppressors. *Nat Rev Cancer* **16,** 413–430 (2016).
100. Tronchère, H., Wang, J. & Fu, X. D. A protein related to splicing factor U2AF35 that interacts with U2AF65 and SR proteins in splicing of pre-mRNA. *Nature* **388,** 397–400 (1997).
101. Madan, V. *et al.* Aberrant splicing of U12-type introns is the hallmark of ZRSR2 mutant myelodysplastic syndrome. *Nat Commun* **6,** 6042 (2015).
102. Grubach, L. *et al.* Gene expression profiling of Polycomb, Hox and Meis genes in patients with acute myeloid leukaemia. *Eur. J. Haematol.* **81,** 112–122 (2008).
103. Mills, K. I. *et al.* Microarray-based classifiers and prognosis models identify subgroups with distinct clinical outcomes and high risk of AML transformation of myelodysplastic syndrome. *Blood* **114,** 1063–1072 (2009).
104. Haferlach, T. *et al.* Clinical utility of microarray-based gene expression profiling in the diagnosis and subclassification of leukemia: report from the International Microarray Innovations in Leukemia Study Group. *J. Clin. Oncol.* **28,** 2529–2537 (2010).
105. Visani, G. *et al.* Gene expression profile predicts response to the combination of tosedostat and low-dose cytarabine in elderly AML. *Blood Adv* **4,** 5040–5049 (2020).
106. Adamia, S. *et al.* A genome-wide aberrant RNA splicing in patients with acute myeloid leukemia identifies novel potential disease markers and therapeutic targets. *Clin. Cancer Res.* **20,** 1135–1145 (2014).

107. Yang, Y.-T. *et al.* The prognostic significance of global aberrant alternative splicing in patients with myelodysplastic syndrome. *Blood Cancer J* **8,** 1–11 (2018).
108. Thorsteinsdottir, U., Kroon, E., Jerome, L., Blasi, F. & Sauvageau, G. Defining Roles for HOX and MEIS1 Genes in Induction of Acute Myeloid Leukemia. *Mol. Cell. Biol.* **21,** 224–234 (2001).
109. Kroon, E., Thorsteinsdottir, U., Mayotte, N., Nakamura, T. & Sauvageau, G. NUP98-HOXA9 expression in hemopoietic stem cells induces chronic and acute myeloid leukemias in mice. *The EMBO Journal* **20,** 350–361 (2001).
110. Sauvageau, G. *et al.* Overexpression of HOXB4 in hematopoietic cells causes the selective expansion of more primitive populations in vitro and in vivo. *Genes Dev.* **9,** 1753–1765 (1995).
111. Mullighan, C. G. *et al.* Pediatric acute myeloid leukemia with NPM1 mutations is characterized by a gene expression profile with dysregulated HOX gene expression distinct from MLL-rearranged leukemias. *Leukemia* **21,** 2000–2009 (2007).
112. Stadler, C. R. *et al.* The leukemogenicity of Hoxa9 depends on alternative splicing. *Leukemia* **28,** 1838–1843 (2014).
113. Adamia, S. *et al.* NOTCH2 and FLT3 gene mis-splicings are common events in patients with acute myeloid leukemia (AML): new potential targets in AML. *Blood* **123,** 2816–2825 (2014).
114. Moore, A. S. *et al.* BIRC5 (survivin) splice variant expression correlates with refractory disease and poor outcome in pediatric acute myeloid leukemia: a report from the Children's Oncology Group. *Pediatr Blood Cancer* **61,** 647–652 (2014).
115. Wagner, M. *et al.* In vivo expression of survivin and its splice variant survivin-2B: impact on clinical outcome in acute myeloid leukemia. *Int. J. Cancer* **119,** 1291–1297 (2006).
116. Tanaka, T. *et al.* An acute myeloid leukemia gene, AML1, regulates hemopoietic myeloid cell differentiation and transcriptional activation antagonistically by two alternative spliced forms. *The EMBO Journal* **14,** 341–350 (1995).
117. Negi, A. RNA Splicing Alterations Induce a Cellular Stress Response Associated with Poor Prognosis in Acute Myeloid Leukemia. 1–12 (2020). doi:10.1158/1078-0432.CCR-20-0184
118. Jin, P., Tan, Y., Zhang, W., Li, J. & Wang, K. Prognostic alternative mRNA splicing signatures and associated splicing factors in acute myeloid leukemia. *Neoplasia* **22,** 447–457 (2020).
119. Pabst, C. *et al.* GPR56 identifies primary human acute myeloid leukemia cells with high repopulating potential in vivo. *Blood* **127,** 2018–2027 (2016).
120. Shiozawa, Y. *et al.* Aberrant splicing and defective mRNA production induced by somatic spliceosome mutations in myelodysplasia. *Nat Commun* **9,** 1–16 (2018).
121. Shen, H., Zheng, X., Luecke, S. & Green, M. R. The U2AF35-related protein Urp contacts the 3' splice site to promote U12-type intron splicing and the second step of U2-type intron splicing. *Genes Dev.* **24,** 2389–2394 (2010).

122. Chang, J.-W. *et al.* mTOR-regulated U2af1 tandem exon splicing specifies transcriptome features for translational control. *Nucleic Acids Res.* **47,** 10373–10387 (2019).
123. Babushok, D. V., Bessler, M. & Olson, T. S. Genetic predisposition to myelodysplastic syndrome and acute myeloid leukemia in children and young adults. *Leuk Lymphoma* **57,** 520–536 (2016).
124. Langabeer, S. E. *et al.* A novel RUNX1 mutation in a kindred with familial platelet disorder with propensity to acute myeloid leukaemia: male predominance of affected individuals. *Eur. J. Haematol.* **85,** 552–553 (2010).
125. Smith, M. L., Cavenagh, J. D., Lister, T. A. & Fitzgibbon, J. Mutation of CEBPA in familial acute myeloid leukemia. *N. Engl. J. Med.* **351,** 2403–2407 (2004).
126. Hahn, C. N. *et al.* Heritable GATA2 mutations associated with familial myelodysplastic syndrome and acute myeloid leukemia. *Nat. Genet.* **43,** 1012–1017 (2011).
127. Rio-Machin, A. *et al.* The complex genetic landscape of familial MDS and AML reveals pathogenic germline variants. *Nat Commun* **11,** 1044 (2020).
128. Staley, J. P. & Guthrie, C. An RNA switch at the 5' splice site requires ATP and the DEAD box protein Prp28p. *Molecular Cell* **3,** 55–64 (1999).
129. Chang, Y.-F., Imam, J. S. & Wilkinson, M. F. The Nonsense-Mediated Decay RNA Surveillance Pathway. *Annu. Rev. Biochem.* **76,** 51–74 (2007).
130. Kurosaki, T. & Maquat, L. E. Nonsense-mediated mRNA decay in humans at a glance. *J. Cell. Sci.* **129,** 461–467 (2016).
131. Stalder, L. & Mühlemann, O. The meaning of nonsense. *Trends Cell Biol* **18,** 315–321 (2008).
132. Lewis, B. P., Green, R. E. & Brenner, S. E. Evidence for the widespread coupling of alternative splicing and nonsense-mediated mRNA decay in humans. *Proc Natl Acad Sci USA* **100,** 189–192 (2003).
133. Tange, T. Ø., Shibuya, T., Jurica, M. S. & Moore, M. J. Biochemical analysis of the EJC reveals two new factors and a stable tetrameric protein core. *RNA* **11,** 1869–1883 (2005).
134. Nicholson, P. *et al.* Nonsense-mediated mRNA decay in human cells: mechanistic insights, functions beyond quality control and the double-life of NMD factors. *Cell Mol Life Sci* **67,** 677–700 (2010).
135. Hug, N. & Caceres, J. F. The RNA Helicase DHX34 Activates NMD by Promoting a Transition from the Surveillance to the Decay-Inducing Complex. *CellReports* **8,** 1845–1856 (2014).
136. Melero, R. *et al.* The RNA helicase DHX34 functions as a scaffold for SMG1-mediated UPF1 phosphorylation. *Nat Commun* **7,** 10585 (2016).
137. Kim, V. N., Kataoka, N. & Dreyfuss, G. Role of the nonsense-mediated decay factor hUpf3 in the splicing-dependent exon-exon junction complex. *Science* **293,** 1832–1836 (2001).

138. Lykke-Andersen, J., Shu, M. D. & Steitz, J. A. Human Upf proteins target an mRNA for nonsense-mediated decay when bound downstream of a termination codon. *Cell* **103,** 1121–1131 (2000).
139. Yepiskoposyan, H., Aeschimann, F., Nilsson, D., Okoniewski, M. & Mühlemann, O. Autoregulation of the nonsense-mediated mRNA decay pathway in human cells. *RNA* **17,** 2108–2118 (2011).
140. Mathew, C. G. Fanconi anaemia genes and susceptibility to cancer. *Oncogene* **25,** 5875–5884 (2006).
141. Taniguchi, T. & D'Andrea, A. D. Molecular pathogenesis of Fanconi anemia: recent progress. *Blood* **107,** 4223–4233 (2006).
142. Ozgur, S. *et al.* The conformational plasticity of eukaryotic RNA-dependent ATPases. *FEBS J* **282,** 850–863 (2015).
143. Robert-Paganin, J., Réty, S. & Leulliot, N. Regulation of DEAH/RHA helicases by G-patch proteins. *BioMed Research International* **2015,** 931857 (2015).
144. Colombo, M., Karousis, E. D., Bourquin, J., Bruggmann, R. & Mühlemann, O. Transcriptome-wide identification of NMD-targeted human mRNAs reveals extensive redundancy between SMG6- and SMG7-mediated degradation pathways. *RNA* **23,** 189–201 (2017).
145. Rodríguez, D. *et al.* Mutations in CHD2 cause defective association with active chromatin in chronic lymphocytic leukemia. *Blood* **126,** 195–202 (2015).
146. Sun, X. *et al.* PARP6 acts as an oncogene and positively regulates Survivin in gastric cancer. *RNA*
147. Cloutier, P., Lavallee-Adam, M., Faubert, D., Blanchette, M. & Coulombe, B. Methylation of the DNA/RNA-binding protein Kin17 by METTL22 affects its association with chromatin. *J Proteomics* **100,** 115–124 (2014).
148. Seervai, R. N. H. *et al.* An actin-WHAMM interaction linking SETD2 and autophagy. *Biochemical and Biophysical Research Communications* **558,** 202–208 (2021).
149. Lynch, K. W. & Weiss, A. A model system for activation-induced alternative splicing of CD45 pre-mRNA in T cells implicates protein kinase C and Ras. *Mol. Cell. Biol.* **20,** 70–80 (2000).
150. Yamashita, A. Role of SMG-1-mediated Upf1 phosphorylation in mammalian nonsense-mediated mRNA decay. *Genes Cells* **18,** 161–175 (2013).
151. Bono, F. *et al.* Molecular insights into the interaction of PYM with the Mago-Y14 core of the exon junction complex. *EMBO Rep* **5,** 304–310 (2004).
152. Fiorini, F., Bagchi, D., Le Hir, H. & Croquette, V. Human Upf1 is a highly processive RNA helicase and translocase with RNP remodelling activities. *Nat Commun* **6,** 7581 (2015).
153. Shum, E. Y. *et al.* The Antagonistic Gene Paralogs Upf3a and Upf3b Govern Nonsense-Mediated RNA Decay. *Cell* **165,** 382–395 (2016).
154. Gilliland, D. G. & Tallman, M. S. Focus on acute leukemias. *Cancer Cell* **1,** 417–420 (2002).

155. Kelly, L. M. & Gilliland, D. G. Genetics of myeloid leukemias. *Annu Rev Genomics Hum Genet* **3,** 179–198 (2002).
156. Mort, M., Ivanov, D., Cooper, D. N. & Chuzhanova, N. A. A meta-analysis of nonsense mutations causing human genetic disease. *Hum Mutat* **29,** 1037–1047 (2008).
157. Hwang, J. & Maquat, L. E. Nonsense-mediated mRNA decay (NMD) in animal embryogenesis: to die or not to die, that is the question. *Current Opinion in Genetics & Development* **21,** 422–430 (2011).
158. Vicente-Crespo, M. & Palacios, I. M. Nonsense-mediated mRNA decay and development: shoot the messenger to survive? *Biochem Soc Trans* **38,** 1500–1505 (2010).
159. Nelson, J. O., Moore, K. A., Chapin, A., Elife, J. H.2016. Degradation of Gadd45 mRNA by nonsense-mediated decay is essential for viability. *elifesciences.org* doi:10.7554/eLife.12876.001
160. Jia, J. *et al.* Caspases shutdown nonsense-mediated mRNA decay during apoptosis. *Cell Death Differ*. **22,** 1754–1763 (2015).
161. Wallace, J. A. & O'Connell, R. M. MicroRNAs and acute myeloid leukemia: therapeutic implications and emerging concepts. *Blood* **130,** 1290–1301 (2017).
162. Fabbri, M. *et al.* MicroRNAs and noncoding RNAs in hematological malignancies: molecular, clinical and therapeutic implications. *Leukemia* **22,** 1095–1105 (2008).
163. Starczynowski, D. T. *et al.* Genome-wide identification of human microRNAs located in leukemia-associated genomic alterations. *Blood* **117,** 595–607 (2011).
164. Brumatti, G., Lalaoui, N., Wei, A. H. & Silke, J. 'Did He Who Made the Lamb Make Thee?' New Developments in Treating the "Fearful Symmetry" of Acute Myeloid Leukemia. *Trends in Molecular Medicine* **23,** 264–281 (2017).
165. Bradstock, K. F. *et al.* A randomized trial of high-versus conventional-dose cytarabine in consolidation chemotherapy for adult de novo acute myeloid leukemia in first remission after induction therapy containing high-dose cytarabine. *Blood* **105,** 481–488 (2005).
166. Escobar-Hoyos, L. F. *et al.* Altered RNA Splicing by Mutant p53 Activates Oncogenic RAS Signaling in Pancreatic Cancer. *Cancer Cell* **38,** 198–211.e8 (2020).
167. Martinez, N. M. *et al.* Widespread JNK-dependent alternative splicing induces a positive feedback loop through CELF2-mediated regulation of MKK7 during T-cell activation. *Genes Dev.* **29,** 2054–2066 (2015).