UNIVERSITY OF CALIFORNIA

SANTA CRUZ

Investigating cancer-associated pre-mRNA splicing alterations using short and long-read sequencing technologies

A dissertation submitted in partial
satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

MOLECULAR, CELL AND DEVELOPMENTAL BIOLOGY

by

Cameron M. Soulette

June 2020

The Dissertation of Cameron Soulette
is approved:

Professor Angela N. Brooks, chair

Professor Manuel Ares, Jr

Professor Jeremy R. Sanford

Quentin Williams
Acting Vice Provost and Dean of Graduate Studies

Contents

Figure List…………………………………………………….………….. v

Abstract…………..…………………….…………………………….... vi

Acknowledgements………………………………………………………. viii

Chapter 1 - Introduction……………………………………………….. 1

1.1 - RNA processing in humans………………………………..... 1

1.2 - Alternative-splicing promotes transcriptome diversity…….. 1

1.3 - Cis-elements and trans-splicing factors are targeted in cancer…………………………………………….. 2

1.4 - Characterizing full-length isoform-level splicing alterations …………………………………………….... 4

1.5 - References ………………………………………………… 9

Chapter 2 - Deep intronic mutations alter pre-mRNA splicing across cancer types…………………………………………..… 16

2.1 - Introduction………………………….……………….... 16

2.2 - Results…………………………………………………. 18

2.3 - Discussion………………………..……………………. 22

2.4 - Methods………………………………..…………………. 24

2.5 - Data access……………………………..………………... 25

2.6 - References……………………………………………………. 31

Chapter 3: Nanopore sequencing reveals *U2AF1 S34F*-associated full-length isoforms……………………………………………. 33

3.1 - Introduction…………………………….……………….... 34

3.2 - Results…………………………………………………….... 38

3.3 - Discussion…………………………………………………. 52

3.4 - Methods…………………..……………………………….. 56

3.5 - Figures …………………………………………………………….... 62

3.6 - Acknowledgements………………………………………………... 68

3.7 - Data access ……………………………………………………….. 69

3.8 - References…………………………………………………………. 69

3.9 - Supplemental Information ………………………………………... 70

Figure List

<u>Chapter 2</u>

Figure 1. Workflow and intronic mutation landscape

Figure 2. Mutational sensitivity of intronic sites to splicing alterations

Figure 3. Putative driver mutations alter splicing in oncogenes and tumor suppressor genes

Figure 4. Splicing alteration recurrence

<u>Chapter 3</u>

Figure 1. Full-length isoform sequencing analysis and workflow

Figure 2. FLAIR captures HBE3kt transcriptome complexity

Figure 3. Nanopore data recapitulates *U2AF1 S34F* splicing signature

Figure 4. U2AF1 S34F-associated isoform usage changes

Figure 5. S34F-associated full-length isoform expression alterations

Figure 6. U2AF1 S34F induces shifts in isoform productivity

Abstract

Investigating cancer-associated pre-mRNA splicing alterations using short and long-read sequencing technologies

by
Cameron M. Soulette

Pre-mRNA splicing is a highly regulated step during gene expression and has been shown to be commonly altered across cancers. The basis for splicing alterations and the functional importance of cancer-associated spliced products remain largely unexplored. The scope of this work aims to better understand the basis for cancer-associated splicing alterations and their functional importance.

We first focus on establishing the genetic basis for cancer-associated splicing alterations. As part of the Pan Cancer Analysis of Whole Genomes (PCAWG) consortium, we demonstrate the impact of non-coding intronic mutations by using matched whole-genome and RNA-sequencing data across 1,209 primary tumor samples spanning 27 cancer types. We identify intronic sites beyond canonical acceptor and donor dinucleotides that are sensitive to mutations, including the branchpoint consensus sequences, which is typically missed in exome sequencing based tumor genotyping. We identify tumor suppressor genes and oncogenes with intronic mutations associated with substantial changes in splicing, and identify previously described alterations in the oncogene *EZH2*, as well as uncharacterized changes in oncogenes *MET* and *HRAS*. Altogether, this work provides the first estimates of the extent to which intronic mutations missed by exome-based genotyping contribute to splicing changes in cancer.

The second half of my work reveals the fate and function of spliced products associated with lung adenocarcinoma mutations in the splicing factor *U2AF1.* We conduct high-throughput long-read cDNA sequencing in isogenic human bronchial epithelial cells with and without *U2AF1 S34F* mutation. We demonstrate the utility of our long-read approach for transcriptome studies by identifying 49,366 novel isoforms exclusive to our approach. We show that our long-read data is robust for capturing mutant *U2AF1*-associated transcriptome alterations by comparing event-level alternative splicing changes with a short-read approach. We identify isoform-level expression changes in 198 isoforms, including a novel lncRNA, and immune-related genes. Last, we hypothesize a mechanism by which *U2AF1 S34F* alters translational control of genes through modulating isoform diversity.

Acknowledgements

I would like to thank my advisor Angela Brooks and my thesis committee, as well as my colleagues in the Brooks lab for their collaborative efforts and unwavering support. I’d like to acknowledge and thank all Pan Cancer Analysis of Whole Genomes working groups, with special thanks to PCAWG-working group 3 who organized data wrangling, processing and quality control of all transcriptome data. Special thanks to Gunnar Rasch, Andre Kahles, and Natalie Davison from the PCAWG-working group 3 for their collaborative efforts and thoughtful discussions. I’d also like to acknowledge the UCSC Nanopore Group, including Mark Akeson, Hugh Olsen, Miten Jain, and Chris Vollmers for all of their time, effort, and input related to nanopore sequencing analyses. Thanks to Dennis Fei and Harold Varmus for cell lines used in this work, and their support and guidance on working with HBEC3kt cells. This work could not have been done without the support of UCSC staff, faculty and facilities, with special acknowledgements to Teel Lopez and Grace Kistler-Fair. I’d also like to thank the UCSC faculty and staff.

Chapter 1 - Introduction

1.1 - RNA processing

The flow of genetic information begins with transcription of DNA into RNA, which is mediated through a multimeric protein complex called RNA polymerase. During transcription, RNA polymerase II plays an important role in facilitating the synthesis and processing of nascent premature RNA to mature RNA. The three main RNA processing steps are 5' capping of RNA with a methyl guanosine residue, removal of spliceosomal introns, and 3' cleavage and polyadenylation (McCracken et al. 1997; Cho, Takagi, and Moore 1997; Hirose and Manley 1998; Misteli and Spector 1999; Hirose, Tacke, and Manley 1999). Of these three essential steps in gene expression, splicing is considered the most highly regulated and dynamic process. During splicing, intervening intragenic regions called introns are removed from pre-mRNA and exonic sequence is ligated together by a careful orchestration of small nuclear ribonucleoproteins splicing factors (Wahl, Will, and Lührmann 2009; Will and Lührmann 2011). Over the last decades, regulation of splicing has been shown to be critically important in human health.

1.2 - Alternative-splicing promotes transcriptome diversity

Analysis of high-throughput transcriptome data revealed that more than 95% of human genes undergo a process called alternative splicing, in which a set of exons from the same pre-mRNA are spliced together in unique combinations to produce distinct mature mRNA products (Modrek and Lee 2002; Pan et al. 2008). Early hypotheses predating the observation of alternative splicing as a transcriptome-wide process postulated that it could diversify the proteome (Gilbert 1978), and indeed, studies have shown alternative splicing as a mechanism of proteome diversification (Graveley 2001; Nilsen and Graveley 2010; Irimia and Roy 2014). Moreover, alternative splicing has been shown to modulate other post-transcriptional processes such as translational control, localization, mRNA stability and turnover (Lewis, Green, and Brenner 2003; Ghosh, Stewart, and Matlashewski 2004; Ni et al. 2007; Sterne-Weiler et al. 2013; Floor and Doudna 2016). The many roles described for alternative splicing in modulating mRNA fate seem to be an essential mechanism for fine-tuning transcriptomes, yet dysregulation of alternative splicing is commonly observed in disease.

1.3 - Cis-elements and trans-splicing factors are targeted in cancer

Pivotal work over the last decade has shown that splicing is largely altered across cancers (as discussed in Venables 2004; Venables et al. 2009; ; David and Manley 2010; Oltean and Bates 2014; Sveen et al. 2016; Dvinge et al. 2019; Escobar-Hoyos, Knorr, and Abdel-Wahab 2019). Trans-acting splicing factor proteins are known to be subject to recurrent mutations across different cancer types. In blood malignancies, splicing factors such as *SF3B1* and *U2AF1* are mutated in nearly half of cancer cohorts, suggesting important roles for splicing factor mutants in cancer (Seiler et al. 2018). *U2AF1* mutations in lung cancer are of particular interest, since the functional importance of mutant U2AF1 is unclear. *U2AF1* is an essential splicing factor that functions in the early steps of pre-mRNA splicing to identify the 3’ end of introns

(Krämer and Utans 1991; Berglund, Abovich, and Rosbash 1998). In lung adenocarcinomas, mutant U2AF1 has been shown to have an altered binding affinity to its pre-mRNA, and lead to widespread splicing alterations (Przychodzen et al. 2013; Brooks et al. 2014; Coulon et al. 2014; Ilagan et al. 2015; Shirai et al. 2015; Park et al. 2016; Yip et al. 2017; Palangat et al. 2019; Smith et al. 2019). Although the contribution of mutant U2AF1 has been shown to impact distinct pathways (Park et al. 2016; Palangat et al. 2019; Smith et al. 2019), the importance of its associated spliced variants, and splice variants from other recurrently mutated splicing factor proteins are poorly understood.

Splicing alterations have also been shown to be caused by somatic mutations that disrupt splicing information-rich intronic elements. *MET*, a tyrosine receptor kinase, is an example of a proto-oncogene with known splicing alterations. *MET* was initially identified as a factor that confers drug resistance in lung cancers by activation of the PI3k/Akt pathway via ERBB3, and later discovered that *MET* itself can act as a strong driver (Kawakami et al. 2014; Engelman et al. 2007). In a further investigation of a large-scale lung cancer cohort, recurrent mutations were identified at the intron-exon boundary of *MET* exon 14, causing exclusion of exon 14 during splicing (Onozato et al. 2009; Schrock et al. 2016; Cancer Genome Atlas Research Network 2014). Functional characterization of exon 14 revealed a conserved ubiquitination domain, in which skipping prevents *MET* from undergoing degradation and causes overexpression and acts as a strong oncogenic driver. Mutations in cis-elements such as the ones that occur in *MET* are not an isolated case, and many other examples of tumor suppressor and oncogenes exist (Kahles et al. 2018; Yoshida et al. 2011; Jung et al. 2015).

A large majority of mutations that alter splicing have been shown to have a genetic basis in cis-regulatory elements or the trans factors that act upon them (Jung et al. 2015; Onozato et al. 2009; Takahashi et al. 1990; Ilagan et al. 2015; Schrock et al. 2016; Campbell et al. 2016; Cancer Genome Atlas Research Network 2014). However, some splicing alterations have been shown to occur even in the absence of any observable mutations (Dvinge and Bradley 2015). One possibility that could explain such *ex nihilo* splicing alterations is that the sequencing tools and strategies we use to characterize tumor genomes lack the resolution to identify the causative mutation. Mutations deep within introns, ones typically missed by standard exome-based tumor genotyping, could impact inclusion of adjacent exons. In the second chapter of this work, we explore this idea by analyzing whole genome and matched transcriptome sequencing data from primary tumor samples to determine the extent to which deep intronic mutations impact splicing.

1.4 - Characterizing full-length isoform-level splicing alterations

Our power to identify somatic mutations that alter splicing has rapidly evolved over the last decade. Without question, the advent of next-generation high-throughput sequencing revolutionized cancer genomics and transcriptomics. However, the massive influx of identified spliced variants from next-generation sequencing has left us at a bottleneck, in which the identification of splicing alterations substantially exceeds our power to narrow down, identify, and functionally test key dysregulated events. One avenue for facilitating the characterization of key events could be accomplished by increasing the resolution at which we identify disease associated RNA molecules.

The throughput and accuracy of short-read sequencing makes it a powerful tool for identifying distinct mRNA processing events, such as individual splicing events. Short-read sequencing is indeed useful for reconstructing a mosaic of the original mRNA transcript, but piecing together transcripts that are several kilobases from small 100-250 nucleotide cDNA fragments is computationally challenging. Previous analyses have demonstrated the limitations in short-read transcriptome assembly by benchmarking various assembly methods using real and synthetic RNA-sequencing datasets (Steijger et al. 2013; Engström et al. 2013). These findings unsurprising given that the fragmentation process in short-read library prep unlinks mRNA processing events, such as capping, splicing, and polyadenylation. Moreover, limitations in isoform assembly can also be explained by the difficulty in resolving alternative-splicing (AS) patterns, which increases the complexity of resolving which events may have occurred on the same mRNA.

Identifying retained introns is another difficult task to accomplish using short-reads. Intron retention (IR) is a type of AS in which removal of intronic sequence is repressed by splicing factors during pre-mRNA processing, imparting various functions such as coding for protein or regulating export and turnover of mRNA (

Galante et al. 2004; Braunschweig et al. 2014). IR events are also of particular interest in cancer because they have been observed to be a widespread phenomenon in solid and blood malignancies, and have been implicated as a mechanism of tumor suppressor inactivation (Simon et al. 2014; Dvinge and Bradley 2015; Jung et al. 2015; Sowalsky et al. 2015; Koh et al. 2015). The difficulty underlying IR characterization is a combination of intron length and loss of short-read connectivity. In a recent re-analysis of colorectal RNA-seq data from The Cancer Genome Atlas, investigators revealed extensive misclassification of IR events, mostly causes by read coverage in overlapping transcript features and non-uniform coverage across introns (Wang and Rio 2018). The solution for more accurate IR and isoform detection has been proposed and proven (Mercer et. al. 2012), longer sequenced reads containing the complete set of mRNA processing decisions can lead to better characterization.

Emerging long-read sequencing technologies, commonly referred to as third-generation sequencing, can help resolve complex AS patterns such as IR events. Long-read technologies can sequence cDNA and RNA molecules that range in size between a few hundred bases to several hundred kilobases of sequence. These ultra-long reads are achieved through the sequencing of unfragmented single molecules, which provides new opportunities to potentially characterize full-length reverse-transcribed cDNA molecules, or entire RNAs. Sequencing of an entire molecule potentially captures, in a single sequenced read, all of the components that define an mRNA isoform such as the (i) transcript start, (ii) the entire set of spliced junctions, and (iii) the polyadenylation cleavage site. Moreover, reading of an entire unfragmented molecule maintains exon connectivity, which resolves complex AS patterns, such as IR. There are several long-read sequencing technologies, but the two most established companies are Pacific Biosciences and Oxford Nanopore Technologies. While both of these companies produce technologies capable of sequencing molecules that fall within the same size range, their approach to achieving ultra-long reads are different. Pacific Biosciences's Single Molecule, Real-Time (SMRT) sequencing utilizes the incorporation of fluorescently labeled nucleotides, a sequencing by synthesis approach. In contrast, the Oxford Nanopore Minion uses a sequencing flow cell, in which molecules are electronically sequenced by measuring changes in voltage as RNA or cDNA is passed through nano-sized pores embedded within a matrix membrane.

Long-read technologies have already been used to characterize novel cancer-specific isoforms which have not been found by either first or second-generation sequencing approaches. *BRCA1* is an example of a gene with several spliced isoforms that have been extensively studied in the context of breast cancer (Miki et al. 1994; Xu et al. 1995; Fortin et al. 2005; Sevcik et al. 2012; Dosil, Tosar, and Cañadas 2010). BRCA1 is a recurrently mutated locus that spans a 100 kilobase genomic region, and is composed of 22 coding exons (Miki et al. 1994). In a large-scale effort by the Evidence-based Network for the Interpretation of Germ-line Mutant Alleles (ENIGMA) consortium (Spurdle et al. 2012), authors utilized various data from first and second generation sequencing methods to characterize 63 *BRCA1* spliced isoforms of clinical significance (Colombo et al. 2014). Although this large-scale effort produced a seemingly complete catalogue of *BRCA1* isoforms, recent long-read sequencing revealed 20 novel isoforms missed by previous analyses, 18 of which contain co-occurring AS patterns (de Jong et al. 2017). The functional importance of these spliced isoforms have not been tested, but isoforms with AS patterns affecting RING and BRCT domains of BRCA1, domains that involved in BCRA1 nuclear localization and DNA repair activity, are hypothesized to be clinically relevant. This particular example highlights the added benefit that long-read sequencing provides over traditional sequencing approaches.

The work presented in my third chapter uses a long-read sequencing approach to characterize isoform-specific changes in response to *U2AF1* mutations observed in lung adenocarcinoma. We develop a platform for full-length isoform discovery to identify cancer-associated isoforms, which will aid in subsequent functional characterization. Our results identify a number of isoform-specific changes caused by changes in splicing, polyadenylation, transcription start site usage, and overall expression. We find an enrichment of premature termination codon-containing isoforms downregulated at the level of expression, and use polysome profiling data to show that affected isoforms also have significant changes in their association with polysome fractions.

1.5 - References

Adams, M. D., S. E. Celniker, R. A. Holt, C. A. Evans, J. D. Gocayne, P. G. Amanatides, S. E. Scherer, et al. 2000. “The Genome Sequence of Drosophila Melanogaster.” *Science* 287 (5461): 2185–95.

Berglund, J. A., N. Abovich, and M. Rosbash. 1998. “A Cooperative Interaction between U2AF65 and mBBP/SF1 Facilitates Branchpoint Region Recognition.” *Genes & Development* 12 (6): 858–67.

Braunschweig, Ulrich, Nuno L. Barbosa-Morais, Qun Pan, Emil N. Nachman, Babak Alipanahi, Thomas Gonatopoulos-Pournatzis, Brendan Frey, Manuel Irimia, and Benjamin J. Blencowe. 2014. “Widespread Intron Retention in Mammals Functionally Tunes Transcriptomes.” *Genome Research* 24 (11): 1774–86.

Brooks, Angela N., Peter S. Choi, Luc de Waal, Tanaz Sharifnia, Marcin Imielinski, Gordon Saksena, Chandra Sekhar Pedamallu, et al. 2014. “A Pan-Cancer Analysis of Transcriptome Changes Associated with Somatic Mutations in U2AF1 Reveals Commonly Altered Splicing Events.” *PloS One* 9 (1): e87361.

Campbell, Joshua D., Anton Alexandrov, Jaegil Kim, Jeremiah Wala, Alice H. Berger, Chandra Sekhar Pedamallu, Sachet A. Shukla, et al. 2016. “Distinct Patterns of Somatic Genome Alterations in Lung Adenocarcinomas and Squamous Cell Carcinomas.” *Nature Genetics* 48 (6): 607–16.

Cancer Genome Atlas Research Network. 2014. “Comprehensive Molecular Profiling of Lung Adenocarcinoma.” *Nature* 511 (7511): 543–50.

C. elegans Sequencing Consortium. 1998. “Genome Sequence of the Nematode C. Elegans: A Platform for Investigating Biology.” *Science* 282 (5396): 2012–18.

Cho, E. J., T. Takagi, and C. R. Moore. 1997. “mRNA Capping Enzyme Is Recruited to the Transcription Complex by Phosphorylation of the RNA Polymerase II Carboxy-Terminal Domain.” *Genes*. http://genesdev.cshlp.org/content/11/24/3319.short.

Clark, Serena L., Ana M. Rodriguez, Russell R. Snyder, Gary D. V. Hankins, and Darren Boehning. 2012. “Structure-Function Of The Tumor Suppressor BRCA1.” *Computational and Structural Biotechnology Journal* 1 (1). https://doi.org/10.5936/csbj.201204005.

Colombo, Mara, Marinus J. Blok, Phillip Whiley, Marta Santamariña, Sara Gutiérrez-Enríquez, Atocha Romero, Pilar Garre, et al. 2014. “Comprehensive Annotation of Splice Junctions Supports Pervasive Alternative Splicing at the BRCA1 Locus: A

Report from the ENIGMA Consortium.” *Human Molecular Genetics* 23 (14): 3666–80.

Coulon, Antoine, Matthew L. Ferguson, Valeria de Turris, Murali Palangat, Carson C. Chow, and Daniel R. Larson. 2014. “Kinetic Competition during the Transcription Cycle Results in Stochastic RNA Processing.” *eLife* 3 (October). https://doi.org/10.7554/eLife.03939.

David, Charles J., and James L. Manley. 2010. “Alternative Pre-mRNA Splicing Regulation in Cancer: Pathways and Programs Unhinged.” *Genes & Development* 24 (21): 2343–64.

Dosil, V., A. Tosar, and C. Cañadas. 2010. “Alternative Splicing and Molecular Characterization of Splice Site Variants: BRCA1 C. 591C> T as a Case Study.” *Clinical*. http://clinchem.aaccjnls.org/content/56/1/53.short.

Dvinge, Heidi, and Robert K. Bradley. 2015. “Widespread Intron Retention Diversifies Most Cancer Transcriptomes.” *Genome Medicine* 7 (1): 45.

Dvinge, Heidi, Jamie Guenthoer, Peggy L. Porter, and Robert K. Bradley. 2019. “RNA Components of the Spliceosome Regulate Tissue- and Cancer-Specific Alternative Splicing.” *Genome Research* 29 (10): 1591–1604.

Engelman, Jeffrey A., Kreshnik Zejnullahu, Tetsuya Mitsudomi, Youngchul Song, Courtney Hyland, Joon Oh Park, Neal Lindeman, et al. 2007. “MET Amplification Leads to Gefitinib Resistance in Lung Cancer by Activating ERBB3 Signaling.” *Science* 316 (5827): 1039–43.

Engström, Pär G., Tamara Steijger, Botond Sipos, Gregory R. Grant, André Kahles, Gunnar Rätsch, Nick Goldman, et al. 2013. “Systematic Evaluation of Spliced Alignment Programs for RNA-Seq Data.” *Nature Methods* 10 (12): 1185–91.

Escobar-Hoyos, Luisa, Katherine Knorr, and Omar Abdel-Wahab. 2019. “Aberrant RNA Splicing in Cancer.” *Annual Review of Cancer Biology* 3 (1): 167–85.

Fetzer, S., Tworek, H. A., Piver, M. S., & Dicioccio, R. A. (1998). An alternative splice site junction in exon 1a of the BRCA1 gene. *Cancer genetics and cytogenetics*, *105*(1), 90-92.

Floor, Stephen N., and Jennifer A. Doudna. 2016. “Tunable Protein Synthesis by Transcript Isoforms in Human Cells.” *eLife* 5 (January). https://doi.org/10.7554/eLife.10921.

Fortin, Jessyka, Anne-Marie Moisan, Martine Dumont, Gilles Leblanc, Yvan Labrie, Francine Durocher, Paul Bessette, et al. 2005. “A New Alternative Splice Variant of BRCA1 Containing an Additional in-Frame Exon.” *Biochimica et Biophysica Acta* 1731 (1): 57–65.

Galante, Pedro Alexandre Favoretto, Noboru Jo Sakabe, Natanja Kirschbaum-Slager, and Sandro José de Souza. 2004. “Detection and Evaluation of Intron Retention Events in the Human Transcriptome.” *RNA* 10 (5): 757–65.

Ghosh, Anirban, Deborah Stewart, and Greg Matlashewski. 2004. “Regulation of Human p53 Activity and Cell Localization by Alternative Splicing.” *Molecular and Cellular Biology* 24 (18): 7987–97.

Gilbert, W. 1978. “Why Genes in Pieces?” *Nature* 271 (5645): 501.

Graveley, B. R. 2001. “Alternative Splicing: Increasing Diversity in the Proteomic World.” *Trends in Genetics: TIG* 17 (2): 100–107.

Hirose, Y., and J. L. Manley. 1998. “RNA Polymerase II Is an Essential mRNA Polyadenylation Factor.” *Nature* 395 (6697): 93–96.

Hirose, Y., R. Tacke, and J. L. Manley. 1999. “Phosphorylated RNA Polymerase II Stimulates Pre-mRNA Splicing.” *Genes & Development* 13 (10): 1234–39.

Ilagan, Janine O., Aravind Ramakrishnan, Brian Hayes, Michele E. Murphy, Ahmad S. Zebari, Philip Bradley, and Robert K. Bradley. 2015. “U2AF1 Mutations Alter Splice Site Recognition in Hematological Malignancies.” *Genome Research* 25 (1): 14–26.

Irimia, Manuel, and Scott William Roy. 2014. “Origin of Spliceosomal Introns and Alternative Splicing.” *Cold Spring Harbor Perspectives in Biology* 6 (6). https://doi.org/10.1101/cshperspect.a016071.

Jong, Lucy C. de, Simone Cree, Vanessa Lattimore, George A. R. Wiggins, Amanda B. Spurdle, kConFab Investigators, Allison Miller, Martin A. Kennedy, and Logan C. Walker. 2017. “Nanopore Sequencing of Full-Length BRCA1 mRNA Transcripts Reveals Co-Occurrence of Known Exon Skipping Events.” *Breast Cancer Research: BCR* 19 (1): 127.

Jung, Hyunchul, Donghoon Lee, Jongkeun Lee, Donghyun Park, Yeon Jeong Kim, Woong-Yang Park, Dongwan Hong, Peter J. Park, and Eunjung Lee. 2015. “Intron Retention Is a Widespread Mechanism of Tumor-Suppressor Inactivation.” *Nature Genetics* 47 (11): 1242–48.

Kahles, André, Kjong-Van Lehmann, Nora C. Toussaint, Matthias Hüser, Stefan G. Stark, Timo Sachsenberg, Oliver Stegle, et al. 2018. "Comprehensive Analysis of Alternative Splicing Across Tumors from 8,705 Patients." *Cancer Cell* 34 (2): 211–24.e6.

Kawakami, Hisato, Isamu Okamoto, Wataru Okamoto, Junko Tanizaki, Kazuhiko Nakagawa, and Kazuto Nishio. 2014. "Targeting MET Amplification as a New Oncogenic Driver." *Cancers* 6 (3): 1540–52.

Koh, Cheryl M., Marco Bezzi, Diana H. P. Low, Wei Xia Ang, Shun Xie Teo, Florence P. H. Gay, Muthafar Al-Haddawi, et al. 2015. "MYC Regulates the Core Pre-mRNA Splicing Machinery as an Essential Step in Lymphomagenesis." *Nature* 523 (7558): 96–100.

Krämer, A., and Ulrike Utans. 1991. "Three Protein Factors (SF1, SF3 and U2AF) Function in Pre-Splicing Complex Formation in Addition to snRNPs." *The EMBO Journal* 10 (6): 1503–9.

Lewis, Benjamin P., Richard E. Green, and Steven E. Brenner. 2003. "Evidence for the Widespread Coupling of Alternative Splicing and Nonsense-Mediated mRNA Decay in Humans." *Proceedings of the National Academy of Sciences of the United States of America* 100 (1): 189–92.

McCracken, Susan, Nova Fong, Emanuel Rosonina, Krassimir Yankulov, Greg Brothers, David Siderovski, Andrew Hessel, et al. 1997. "5′-Capping Enzymes Are Targeted to Pre-mRNA by Binding to the Phosphorylated Carboxy-Terminal Domain of RNA Polymerase II." *Genes & Development* 11 (24): 3306–18.

Miki, Y., J. Swensen, D. Shattuck-Eidens, P. A. Futreal, K. Harshman, S. Tavtigian, Q. Liu, C. Cochran, L. M. Bennett, and W. Ding. 1994. "A Strong Candidate for the Breast and Ovarian Cancer Susceptibility Gene BRCA1." *Science* 266 (5182): 66–71.

Misteli, T., and D. L. Spector. 1999. "RNA Polymerase II Targets Pre-mRNA Splicing Factors to Transcription Sites in Vivo." *Molecular Cell* 3 (6): 697–705.

Modrek, Barmak, and Christopher Lee. 2002. "A Genomic View of Alternative Splicing." *Nature Genetics* 30 (1): 13–19.

Ni, Julie Z., Leslie Grate, John Paul Donohue, Christine Preston, Naomi Nobida, Georgeann O'Brien, Lily Shiue, Tyson A. Clark, John E. Blume, and Manuel Ares. 2007. "Ultraconserved Elements Are Associated with Homeostatic Control of Splicing Regulators by Alternative Splicing and Nonsense-Mediated Decay." *Genes & Development* 21 (6): 708–18.

Nilsen, Timothy W., and Brenton R. Graveley. 2010. “Expansion of the Eukaryotic Proteome by Alternative Splicing.” *Nature* 463 (7280): 457–63.

Oltean, S., and D. O. Bates. 2014. “Hallmarks of Alternative Splicing in Cancer.” *Oncogene* 33 (46): 5311–18.

Onozato, Ryoichi, Takayuki Kosaka, Hiroyuki Kuwano, Yoshitaka Sekido, Yasushi Yatabe, and Tetsuya Mitsudomi. 2009. “Activation of MET by Gene Amplification or by Splice Mutations Deleting the Juxtamembrane Domain in Primary Resected Lung Cancers.” *Journal of Thoracic Oncology: Official Publication of the International Association for the Study of Lung Cancer* 4 (1): 5–11.

Palangat, Murali, Dimitrios G. Anastasakis, Dennis Liang Fei, Katherine E. Lindblad, Robert Bradley, Christopher S. Hourigan, Markus Hafner, and Daniel R. Larson. 2019. “The Splicing Factor U2AF1 Contributes to Cancer Progression through a Noncanonical Role in Translation Regulation.” *Genes & Development* 33 (9-10): 482–97.

Pan, Qun, Ofer Shai, Leo J. Lee, Brendan J. Frey, and Benjamin J. Blencowe. 2008. “Deep Surveying of Alternative Splicing Complexity in the Human Transcriptome by High-Throughput Sequencing.” *Nature Genetics* 40 (12): 1413–15.

Park, Sung Mi, Jianhong Ou, Lynn Chamberlain, Tessa M. Simone, Huan Yang, Ching-Man Virbasius, Abdullah M. Ali, et al. 2016. “U2AF35(S34F) Promotes Transformation by Directing Aberrant ATG7 Pre-mRNA 3’ End Formation.” *Molecular Cell* 62 (4): 479–90.

Przychodzen, Bartlomiej, Andres Jerez, Kathryn Guinta, Mikkael A. Sekeres, Richard Padgett, Jaroslaw P. Maciejewski, and Hideki Makishima. 2013. “Patterns of Missplicing due to Somatic U2AF1 Mutations in Myeloid Neoplasms.” *Blood* 122 (6): 999–1006.

Schrock, Alexa B., Garrett M. Frampton, James Suh, Zachary R. Chalmers, Mark Rosenzweig, Rachel L. Erlich, Balazs Halmos, et al. 2016. “Characterization of 298 Patients with Lung Cancer Harboring MET Exon 14 Skipping Alterations.” *Journal of Thoracic Oncology: Official Publication of the International Association for the Study of Lung Cancer* 11 (9): 1493–1502.

Seiler, Michael, Shouyong Peng, Anant A. Agrawal, James Palacino, Teng Teng, Ping Zhu, Peter G. Smith, Cancer Genome Atlas Research Network, Silvia Buonamici, and Lihua Yu. 2018. “Somatic Mutational Landscape of Splicing Factor Genes and Their Functional Consequences across 33 Cancer Types.” *Cell Reports* 23 (1): 282–96.e4.

Sevcik, Jan, Martin Falk, Petra Kleiblova, Filip Lhota, Lenka Stefancikova, Marketa Janatova, Lenka Weiterova, et al. 2012. “The BRCA1 Alternative Splicing Variant Δ14-15 with an in-Frame Deletion of Part of the Regulatory Serine-Containing Domain (SCD) Impairs the DNA Repair Capacity in MCF-7 Cells.” *Cellular Signalling* 24 (5): 1023–30.

Shirai, Cara Lunn, James N. Ley, Brian S. White, Sanghyun Kim, Justin Tibbitts, Jin Shao, Matthew Ndonwi, et al. 2015. “Mutant U2AF1 Expression Alters Hematopoiesis and Pre-mRNA Splicing In Vivo.” *Cancer Cell* 27 (5): 631–43.

Simon, Jeremy M., Kathryn E. Hacker, Darshan Singh, A. Rose Brannon, Joel S. Parker, Matthew Weiser, Thai H. Ho, et al. 2014. “Variation in Chromatin Accessibility in Human Kidney Cancer Links H3K36 Methyltransferase Loss with Widespread RNA Processing Defects.” *Genome Research* 24 (2): 241–50.

Smith, Molly A., Gaurav S. Choudhary, Andrea Pellagatti, Kwangmin Choi, Lyndsey C. Bolanos, Tushar D. Bhagat, Shanisha Gordon-Mitchell, et al. 2019. “U2AF1 Mutations Induce Oncogenic IRAK4 Isoforms and Activate Innate Immune Pathways in Myeloid Malignancies.” *Nature Cell Biology* 21 (5): 640–50.

Sowalsky, Adam G., Zheng Xia, Liguo Wang, Hao Zhao, Shaoyong Chen, Glenn J. Bubley, Steven P. Balk, and Wei Li. 2015. “Whole Transcriptome Sequencing Reveals Extensive Unspliced mRNA in Metastatic Castration-Resistant Prostate Cancer.” *Molecular Cancer Research: MCR* 13 (1): 98–106.

Spurdle, Amanda B., Sue Healey, Andrew Devereau, Frans B. L. Hogervorst, Alvaro N. A. Monteiro, Katherine L. Nathanson, Paolo Radice, et al. 2012. “ENIGMA—Evidence-Based Network for the Interpretation of Germline Mutant Alleles: An International Initiative to Evaluate Risk and Clinical Significance Associated with Sequence Variation in BRCA1 and BRCA2 Genes.” *Human Mutation* 33 (1): 2–7.

Steijger, Tamara, Josep F. Abril, Pär G. Engström, Felix Kokocinski, RGASP Consortium, Tim J. Hubbard, Roderic Guigó, Jennifer Harrow, and Paul Bertone. 2013. “Assessment of Transcript Reconstruction Methods for RNA-Seq.” *Nature Methods* 10 (12): 1177–84.

Sterne-Weiler, Timothy, Rocio Teresa Martinez-Nunez, Jonathan M. Howard, Ivan Cvitovik, Sol Katzman, Muhammad A. Tariq, Nader Pourmand, and Jeremy R. Sanford. 2013. “Frac-Seq Reveals Isoform-Specific Recruitment to Polyribosomes.” *Genome Research* 23 (10): 1615–23.

Sveen, A., S. Kilpinen, A. Ruusulehto, R. A. Lothe, and R. I. Skotheim. 2016. “Aberrant RNA Splicing in Cancer; Expression Changes and Driver Mutations of Splicing Factor Genes.” *Oncogene* 35 (19): 2413–27.

Takahashi, T., D. D'Amico, I. Chiba, D. L. Buchhagen, and J. D. Minna. 1990. "Identification of Intronic Point Mutations as an Alternative Mechanism for p53 Inactivation in Lung Cancer." *The Journal of Clinical Investigation* 86 (1): 363–69.

Venables, Julian P. 2004. "Aberrant and Alternative Splicing in Cancer." *Cancer Research* 64 (21): 7647–54.

Venables, Julian P., Roscoe Klinck, Chushin Koh, Julien Gervais-Bird, Anne Bramard, Lyna Inkel, Mathieu Durand, et al. 2009. "Cancer-Associated Regulation of Alternative Splicing." *Nature Structural & Molecular Biology* 16 (6): 670–76.

Venter, J. C., M. D. Adams, E. W. Myers, P. W. Li, R. J. Mural, G. G. Sutton, H. O. Smith, et al. 2001. "The Sequence of the Human Genome." *Science* 291 (5507): 1304–51.

Wahl, Markus C., Cindy L. Will, and Reinhard Lührmann. 2009. "The Spliceosome: Design Principles of a Dynamic RNP Machine." *Cell* 136 (4): 701–18.

Wang, Qingqing, and Donald C. Rio. 2018. "JUM Is a Computational Method for Comprehensive Annotation-Free Analysis of Alternative Pre-mRNA Splicing Patterns." *Proceedings of the National Academy of Sciences of the United States of America* 115 (35): E8181–90.

Will, Cindy L., and Reinhard Lührmann. 2011. "Spliceosome Structure and Function." *Cold Spring Harbor Perspectives in Biology* 3 (7). https://doi.org/10.1101/cshperspect.a003707.

Xu, C. F., M. A. Brown, J. A. Chambers, B. Griffiths, H. Nicolai, and E. Solomon. 1995. "Distinct Transcription Start Sites Generate Two Forms of BRCA1 mRNA." *Human Molecular Genetics* 4 (12): 2259–64.

Yip, Bon Ham, Violetta Steeples, Emmanouela Repapi, Richard N. Armstrong, Miriam Llorian, Swagata Roy, Jacqueline Shaw, et al. 2017. "The U2AF1S34F Mutation Induces Lineage-Specific Splicing Alterations in Myelodysplastic Syndromes." *The Journal of Clinical Investigation* 127 (9): 3557.

Yoshida, Kenichi, Masashi Sanada, Yuichi Shiraishi, Daniel Nowak, Yasunobu Nagata, Ryo Yamamoto, Yusuke Sato, et al. 2011. "Frequent Pathway Mutations of Splicing Machinery in Myelodysplasia." *Nature* 478 (7367): 64–69.

Chapter 2 - Deep intronic mutations alter pre-mRNA splicing across cancer types

2.1 - Introduction

Tumor transcriptome sequencing has revealed a striking pattern in which splicing is commonly altered across cancer types. The basis for such alterations have been shown to be partially explained by mutations adjacent to exon-intron boundaries (Venables, 2004; Supek *et al.*, 2014; Jung *et al.*, 2015; Shiraishi *et al.*, 2018). In a previous analysis of 1,812 exome sequencing datasets from cancer patients, investigators estimate that >10% of mutations occur at the GT/AG dinucleotides that define the 5' donor and 3' acceptor site disrupt splicing (Jung *et al.*, 2015). Moreover, mutations that alter splicing were shown to produce splice variants with premature termination codons in tumor suppressor genes (TSGs), suggesting an oncogenic role for TSG inactivation in cancers (Jung *et al.*, 2015). Despite these revelations, the extent to which intronic mutations that alter splicing are selected for in cancer have been almost exclusively, except in the case of MET and BRCA1 (Onozato *et al.*, 2009), limited to GT/AG donor and acceptor sites (Takahashi *et al.*, 1990). The importance of mutations beyond GT/AG dinucleotides has not been comprehensively assessed.

Our work here aims to establish the basis for cancer-associated RNA splicing alterations. We use a highly curated dataset composed of 1,209 whole-genomes from primary tumor samples with matched RNA-sequencing to identify intronic mutations that alter splicing. We examine the landscape of intronic mutations by investigating their impact on exon skipping and intron retention. We find sites beyond the GT/AG donor and acceptor site that are sensitive to splicing changes, including parts of the polypyrimidine (pY) tract and branchpoint sequences. We find mutations associated with splicing changes in numerous TSGs, and uncharacterized splicing changes in known driver oncogenes HRAS, MEF2B, MET, PPP2R1A, and MAX. Last, we use a permutation-based approach to identify genes under positive selection for splicing alterations. Altogether, our work provides a comprehensive view on the contribution of non-coding intronic mutations on cancer-associated splicing alterations.

2.2 - Results

Splicing alteration landscape across cancer types

We looked to capture the general landscape of splicing mutations across introns. To do this we utilized available somatic mutation calls from tumor samples with whole-genome sequencing and matched RNA-seq (Figure 1A) (PCAWG Transcriptome Core Group *et al.*, 2018). We overlapped point mutations across exon-intron boundaries to examine the proportion of mutation types across intronic regions (Figure 1C). The most notable feature of mutations at exon-intron boundaries is a qualitative enrichment of mutations directly at the interface between intronic and exonic sequences (positions -1 and +1). The majority of the enrichment is contributed by G>A mutations at the -1 position of the donor and acceptor site, which is likely due to the lack of heterogeneity at these sites, and that transition mutations are more likely than transversions.

Mutations in the extended splice site motif are associated with splicing changes

We next sought to understand the relation between splicing changes and somatic mutations within introns. We first evaluated the impact of intronic mutations on exon skipping events. As observed in previous reports (Jung *et al.*, 2015), we found significant proportions of mutations occurring at the GT/AG donor and acceptor site associated with changes in splicing (66%, 395/590; |z-score|>=3; p-value <0.01 permutation test) (Figure 2A top panel). We examined mutations in the extended splice site motif, positions -1 to -6 for both donor and acceptor sites, and found similar proportions of somatic mutations associated with changes in splicing (50%; 629/1,253; p-value < 0.05, permutation test . The magnitude of nearly all somatic mutations in the extended splicing motif were negative, suggesting that mutations that increase inclusion by enhanced recognition of the donor splice site motif are rare (Figure 2A bottom panel). Consistent with the notion that most splice site mutations are damaging, we found an enrichment of intron retention events for mutations that occurred within the extended splice site motif (Figure 2B). In contrast, a larger proportion of mutations in the donor site were associated with cassette exon splicing changes in comparison to intron retention events (~50% versus ~30%), possibly suggesting that changes in exon skipping may be more permissible than retained introns. Altogether, these results demonstrate that mutations beyond the GT/AG donor and acceptor dinucleotides impact splicing at a significant rate.

Polypyrimidine tracts and branchpoint mutations are enriched for splicing changes

We extended our position-specific mutation analysis to deeper intronic *cis*-regulatory elements. We first looked at mutations harbored in the polypyrimidine (pY) tract, defined as the 35bp region upstream from 3' acceptor sites, and found several sites significantly enriched for changes in exon skipping and retained introns (p-value < 0.05, permutation test) (Figure 2A, B). Given the pyrimidine rich content of pY tracts, we asked if pyrimidine to purine mutations were enriched for splicing changes. We found TTT>TNT and TNT>TGT mutations to be the most enriched for splicing changes (p-value < 0.01; Fisher's exact two-tailed). Disruption of uracil stretches have been shown to be detrimental for splicing (Roscigno, Weiner and Garcia-Blanco, 1993). We next investigated mutations that occur in the branch site consensus sequence using recently published human branchpoint annotation data (Mercer *et al.*, 2015; Signal *et al.*, 2018). We find several sites enriched for splicing changes, including mutations at branchpoint adenosines (p-value <0.05, permutation test) (Figure 2C, D). Although deeper intronic mutations do not impact splicing to the same extent as splice site motif mutations, our data shows that the sequence elements the pY tract and branch site, typically missed by exome-based variant calling, contain sites sensitive to mutations that alter splicing.

Uncharacterized splice site mutations alter known driver genes

We further inspected somatic mutations in the extended splice site motif to determine if any known cancer driver genes are associated with splicing changes. To do this, we filtered for somatic mutations adjacent to skipped exons with substantial splicing changes (|z-score|>=3) in oncogenes and TSGs previously characterized by TUSON (Davoli *et al.*, 2013). Of the 629 somatic mutations in extended splice site motifs associated with changes in splicing, 70 were adjacent to exons in oncogenes and TSGs, 5 and 65 respectively. Only one of the 5 oncogenes, *EZH2*, was identified by the Pan Cancer Analysis of Whole Genomes group aimed toward identifying cancer driver mutations, whereas the rest were considered passenger mutations (Sabarinathan *et al.*, 2017). We used expression data to investigate the magnitude change in splicing for a subset of the putative driver mutations. Mutations at the -1 donor site for oncogenes *HRAS* and *MET* were associated with large changes in percent spliced in (PSI) values (Δ33% relative to cohort average) (Figure 3 left panels). We also found substantial PSI changes (Δ66% & Δ55% relative to cohort average) for mutations at the -3 position relative to TSGs *ARHGAP35* and *RBM10*, both of which were observed in the same sample. We found no evidence for other mutations for *ARHGAP35* and *RBM10,* excluding the possibility that other intronic or exonic mutations could drive skipping at these sites. Although we observed strong changes in splicing for intronic mutations for *ARHGAP35*, *RBM10* and other cancer-relevant genes, very few were characterized as driver mutations. These results demonstrate that intronic sites beyond acceptor and donor dinucleotides are sensitive to mutations and impact splicing.

Tumor suppressor genes are targets for recurrent exon skipping alterations

We decided to use an approach to identify genes with enriched levels of mutations associated with splicing changes. To do this, we used a permutation-based approach that compares the number of mutations associated with splicing changes versus mutations lacking evidence of splicing changes. We reasoned that a gene enriched for mutations that alter splicing should have a large cumulative splicing impact for mutations likely to alter splicing, relative to random intronic mutations (Figure 4A) (Methods). We applied our analysis to our dataset and identified 36 genes with enrichment of splicing alterations (30 with FDR < 0.05, 6 with FDR<0.01) (Figure 4B). We found that 3 of the top genes, TP53, RB1 and RBM10, were classified as

TSG by the TUSON method, and another, POT1, defined as a TSG by COSMIC. In sum, our analysis captures tumor suppressor genes known to be recurrently altered in cancer, and identifies a number of other genes that may be selected for in cancers.

2.3 - Discussion

We conducted a comprehensive analysis of cancer-related intronic mutations and their impact on splicing using whole-genome tumor sequencing. Our work builds on previous studies utilizing whole-exome sequencing and matched RNA-seq by identifying intronic mutations beyond GT/AG donor and acceptor dinucleotides that impact splicing. Namely, we identify extensive enrichment of mutations associated with outlier splicing in the extended splice site donor and acceptor motif for skipped exons and retained introns. In comparable proportions, we identify sites in the pY tract and branch site consensus that are enriched for splicing changes, which are elements that are rarely captured and characterized in exome sequencing-based tumor genotyping.

We conducted a cursory analysis of gene-level recurrent splicing alterations using a permutation-based approach. Consistent with previous reports, we identified tumor suppressor genes amongst the top hits (4/36), supporting the hypothesis of an inactivating role for deep intronic mutations (Jung *et al.*, 2015). We expand on previous results by showing that putative tumor-inactivating mutations are not exclusively enriched in retained introns. Our cohort-based analysis using whole genomes does introduce limitations in the type of alterations we identified, which precludes identification of splicing alterations selected for by the entire cohort. This limitation could be resolved by utilizing matched-normal tissue, yet the number of matched normal samples in this study was too small for such analyses.

Nevertheless, we find a number of mutations associated with changes in splicing cancer-relevant genes. From our mutational landscape analysis (Figure 2), we identified a number of mutations in the donor and acceptor splice site motif that affect tumor suppressor genes and oncogenes. For oncogenes, we identified a somatic mutation at the -1 position of the splice acceptor site of EZH2 exon 19, which was identified by the PCAWG cancer gene driver group (Sabarinathan *et al.*, 2017). In addition to EZH2, we identified 5 additional oncogenes that were not characterized by the PCAWG driver group, including HRAS, MEF2B, MET, PPP2R1A, and MAX. Although the samples with splicing changes in these oncogenes are associated with other identified driver mutations, recent studies suggest that gain and loss of function mutations in strong drivers typically occur in more than 4 genes for a given sample (Sabarinathan *et al.*, 2017). An important next step in these findings is identifying the functional outcome of these associated splicing changes. For example, splicing changes associated with intronic mutations in *MAX* alter the 3’ end of the mRNA, possibly changing the cellular localization of MAX. An interesting hypothesis to explore is the possibility that alternative 3' end processing induced by splice site mutations inadvertently alters MYC/MAX interaction and activity through MAX mis localization. However, splicing changes in MAX and other oncogenes were not found to be recurrent in this study, and it would be more pertinent to establish the recurrence of these splicing alterations on a larger cancer dataset such as The Cancer Genome Atlas.

2.4 - Methods

Processing of somatic mutation and alternative splicing data

Samples and associated metadata were retrieved from the ICGC Data portal (https://dcc.icgc.org). Tumor samples that did not pass whole genome sequencing or RNA-seq quality control assessments (greylist & blacklist samples) were not used for analyses. Alternative splicing quantification was retrieved from

Enrichment of outlier splicing associated with splice sites and branchpoint motifs

We assessed the significance of mutational enrichment for 5' donor and 3' acceptor splice sites, and branch-point intronic regions using a permutation-based approach. Impactful mutations were defined as mutations overlapping exons and introns involved in cassette exon events and intron retention events, in which the PSI-derived z-score was >= 3 or <= -3. For each intronic site, we compared the frequency of observed impactful mutations against frequencies of randomly sampled intronic regions (number of iterations = 10,000). For exonic sites, the null distribution was established from randomly sampled exonic sites. Randomly sampled sites were within a 100-bp window around the 5’ and 3’ splice site. For branch-point regions, the null distribution was drawn from mutations near 3’ acceptor sites (positions -2 to -35) that did not overlap predicted branch site motifs. The p-value was computed as the number of randomly sampled frequencies greater or equal to the observed frequency.

Positive selection of outlier splicing analysis

We developed a permutation-based methodology to detect genes affected by recurrent splicing alterations. We identified genes with mutations that alter splicing rather than genes which are broadly mutated. To do this, we compared the overall impact of mutations expected to alter splicing (near exon-intron boundaries) versus mutations across random intronic regions. For each gene, we measured the overall gene-level change in splicing across samples with mutations near exon-intron boundaries (starting at intronic position -50 to exonic position +5) by taking the average z-score of all mutations within a defined window. We then compared our observed impact against a null distribution which was generated by permuting over all intronic sites for each mutation for each sample. The p value was defined as the number of times a randomly permuted average z-score was >= the observed z-score. P-values were corrected using python scipy package, using the Benjamini-Hochberg method.

2.5 - Figures

FIGURE 1

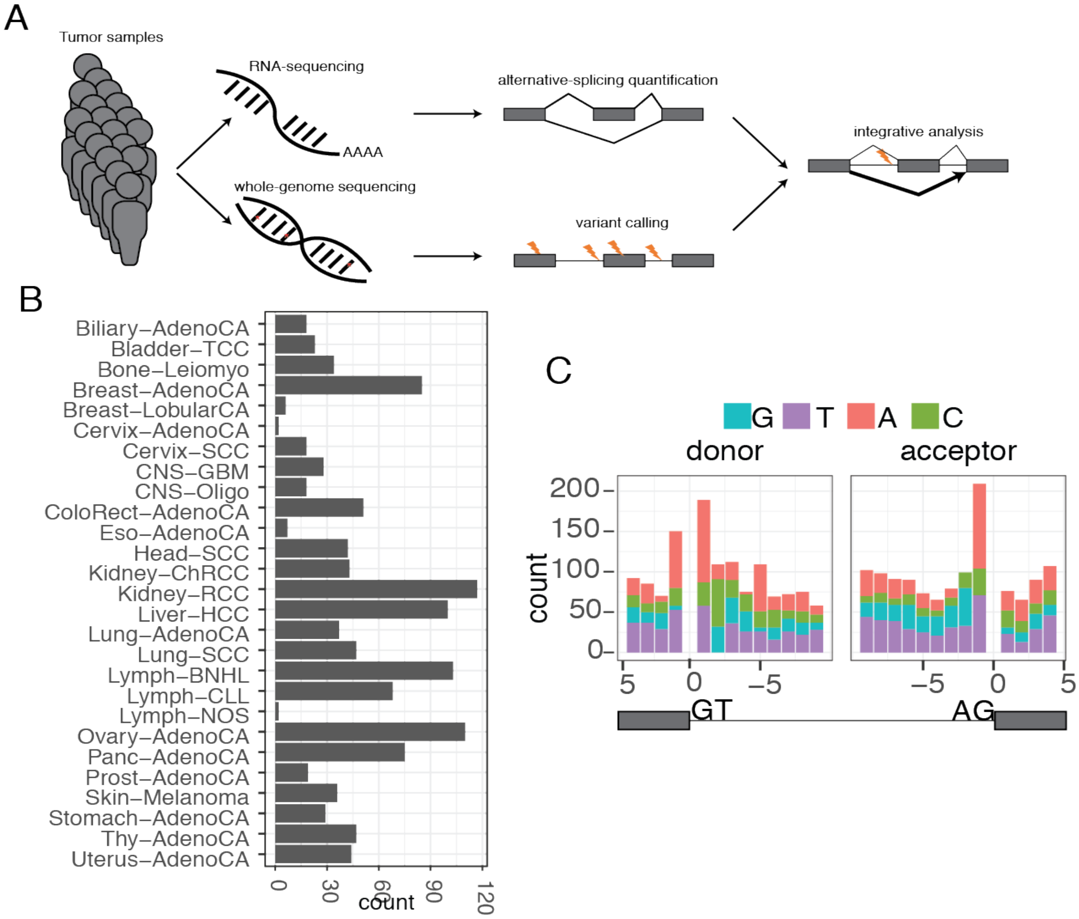

Figure 1 | Workflow and intronic mutation landscape. A) Diagram of general workflow. 1,209 controlled and public access whole genome with matched RNA-seq were variant called by the PCAWG variant calling working group (ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium, 2020). Matched RNA-seq was quantified for alternative splicing by PCAWG transcriptome group (PCAWG Transcriptome Core Group *et al.*, 2018). B) Tissue types and number of samples used in this study. C) Mutational landscape at exon (positions 5 to 0) and intron (positions -1 to -9) boundaries (boundary defined at position 0). Colors denote the type of mutation occurring at each site. Mutations at GT/AG dinucleotide lack GT/AG mutations since all splice sites considered in this analysis conformed to U2 type introns.

FIGURE 2

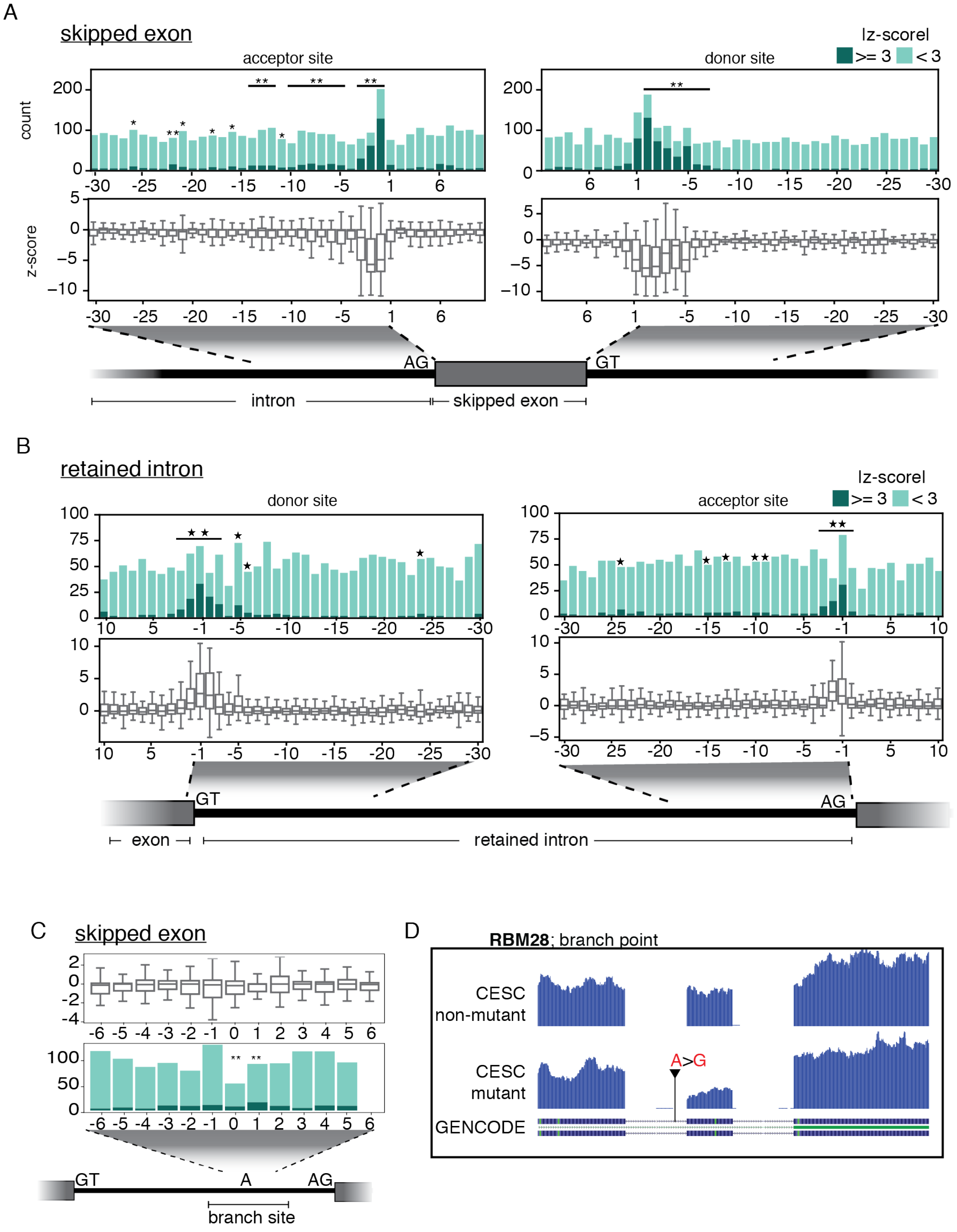

Figure 2 | Mutational sensitivity of intronic sites to splicing alterations. A) Stacked bar plot displays the position of observed somatic mutations overlapping and adjacent to SplAdder skipped exons. Top panel displays total number of mutation, and dark color denotes if the observed mutation occurs near a skipped exon with a |z-score| >= 3. Asterisks denote sites significantly enriched for mutations with |z-score|>=3 by permutation test, * p-value <0.05, **p-value <0.01. Bottom panel displays position of observed somatic mutation overlapping and adjacent to SplAdder skipped exons, and their z-score magnitude. B) Same as (A), but for retained intron events. C) Same as (A), but for branch site regions. D) Example of splicing change due to branch site adenosine mutation. Comparison is a UCSC genome browser shot displaying coverage tracks for tumor and matched normal RNA-seq from the same sample. Red letters denote the nucleotide change.

FIGURE 3

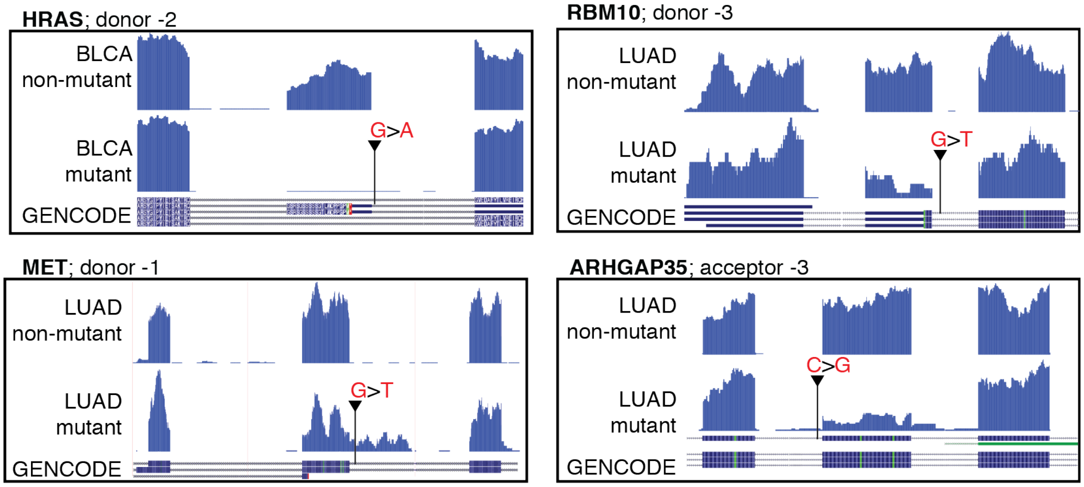

Figure 3 | Putative driver mutations alter splicing in oncogenes and tumor suppressor genes. UCSC Genome browser images displaying coverage tracks for genes with uncharacterized intronic mutations that alter exon skipping events in tumor suppressor and oncogenes. Gene names are labeled above their respective browser image, along with the position of the mutation relative to donor or acceptor sites. GENCODE v19 annotations are displayed below coverage plots. Red letters describe the nucleotide mutation change.

FIGURE 4

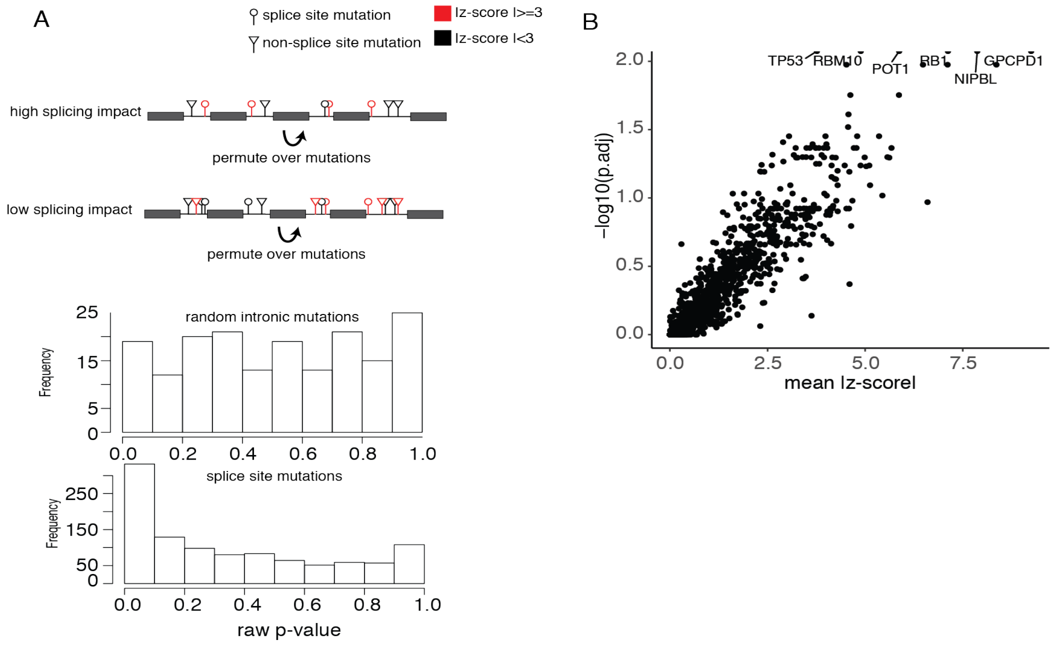

Figure 3 | Splicing alterations are selected for in tumor suppressor genes. A) Schematic of permutation-based approach to identifying genes with enriched splicing alterations. Bottom histograms show raw p-value distributions from analyzing mutations likely to impact splicing (bottom) versus random intronic mutations (top, negative control). B) Volcano plot showing top genes enriched for recurrent splicing alterations.

2.6 - Data Access

All data collected and assembled by the Pan Cancer Analysis of Whole Genomes consortium such as variant calls, patient metadata and alternative splicing quantification is available at https://dcc.icgc.org/releases/PCAWG.

2.7 - References

Davoli, T. *et al.* (2013) ‘Cumulative haploinsufficiency and triplosensitivity drive aneuploidy patterns and shape the cancer genome’, *Cell*. Elsevier, 155(4), pp. 948–962.

ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium (2020) ‘Pan-cancer analysis of whole genomes’, *Nature*, 578(7793), pp. 82–93.

Jung, H. *et al.* (2015) ‘Intron retention is a widespread mechanism of tumor-suppressor inactivation’, *Nature genetics*. nature.com, 47(11), pp. 1242–1248.

Mercer, T. R. *et al.* (2015) ‘Genome-wide discovery of human splicing branchpoints’, *Genome research*, 25(2), pp. 290–303.

Onozato, R. *et al.* (2009) ‘Activation of MET by gene amplification or by splice mutations deleting the juxtamembrane domain in primary resected lung cancers’, *Journal of thoracic oncology: official publication of the International Association for the Study of Lung Cancer*, 4(1), pp. 5–11.

PCAWG Transcriptome Core Group *et al.* (2018) ‘Genomic basis for RNA alterations revealed by whole-genome analyses of 27 cancer types’, *bioRxiv*. doi: 10.1101/183889.

Roscigno, R. F., Weiner, M. and Garcia-Blanco, M. A. (1993) ‘A mutational analysis of the polypyrimidine tract of introns. Effects of sequence differences in pyrimidine tracts on splicing’, *The Journal of biological chemistry*, 268(15), pp. 11222–11229.

Sabarinathan, R. *et al.* (2017) ‘The whole-genome panorama of cancer drivers’, *BioRxiv*. biorxiv.org. Available at: https://www.biorxiv.org/content/10.1101/190330v1.abstract.

Shiraishi, Y. *et al.* (2018) ‘A comprehensive characterization of cis-acting splicing-associated variants in human cancer’, *Genome research*, 28(8), pp. 1111–1125.

Signal, B. *et al.* (2018) ‘Machine learning annotation of human branchpoints’, *Bioinformatics* , 34(6), pp. 920–927.

Supek, F. *et al.* (2014) ‘Synonymous mutations frequently act as driver mutations in human cancers', *Cell*, 156(6), pp. 1324–1335.

Takahashi, T. *et al.* (1990) 'Identification of intronic point mutations as an alternative mechanism for p53 inactivation in lung cancer', *The Journal of clinical investigation*, 86(1), pp. 363–369.

Venables, J. P. (2004) 'Aberrant and alternative splicing in cancer', *Cancer research*. AACR, 64(21), pp. 7647–7654.

Nanopore sequencing reveals *U2AF1 S34F*-associated full-length isoforms

Cameron M. Soulette[1], Eva Hrabeta-Robinson[1], Alison Tang[2], Maximillian G. Marin[2,3,4] , Angela N. Brooks[2*]

[1]Department of Molecular, Cellular & Developmental Biology, [2]Department of Biomolecular Engineering, University of California, Santa Cruz, CA, USA

Present address: [3]Department of Systems Biology, [4]Department of Biomedical Informatics, Harvard Medical School

*Correspondence: anbrooks@ucsc.edu

Running title: U2AF1 S34F-associated full-length isoforms

Keywords: nanopore, U2AF1, splicing factor mutation

Abstract

U2AF1 S34F is one of the most recurrent splicing factor mutations in lung adenocarcinoma (ADC) and has been shown to cause transcriptome-wide pre-mRNA splicing alterations. While *U2AF1 S34F*-associated splicing alterations have been described, the function of altered mRNA isoform changes remains largely unexplored. To better understand the impact *U2AF1 S34F* has on isoform fate and function, we conducted high-throughput long-read cDNA sequencing from isogenic human bronchial epithelial cells with and without *U2AF1 S34F* mutation. We found that nearly 75% (49,366) of our long-read constructed multi-exon isoforms do not overlap GENCODE or short-read assembled isoforms. We found 198 transcript isoforms with significant expression and usage changes caused by *U2AF1 S34F* mutation, including a novel lncRNA. Isoforms from immune-related genes were largely downregulated in mutant cells, none of which were found to have splicing changes. Finally, isoforms likely targeted by nonsense-mediated decay were largely downregulated in *U2AF1 S34F* cells, suggesting that the impact of observed isoform changes may alter the translational output of affected genes. Altogether, long-read sequencing provided additional insights into transcriptome alterations and downstream functional consequences associated with *U2AF1 S34F* mutation.

Introduction

Previous cancer genomic studies across lung adenocarcinoma (ADC) patients have revealed recurrent mutations in the splicing factor *U2AF1 (Brooks et al., 2014; Cancer Genome Atlas Research Network, 2014; Campbell et al., 2016)*. U2AF1 is an essential splicing factor that functions to identify the 3' end of intronic sequence in the early steps of pre-mRNA splicing (Shao *et al.*, 2014). In ADC, the most recurrent *U2AF1* mutation occurs at amino acid residue 34, in which a C>T transition causes a change from serine to phenylalanine (S34F). The impact of *U2AF1 S34F* on pre-mRNA splicing has been widely studied (Przychodzen *et al.*, 2013; Brooks *et al.*, 2014; Coulon *et al.*, 2014; Ilagan *et al.*, 2015; Shirai *et al.*, 2015; Park *et al.*, 2016; Yip *et al.*, 2017; Palangat *et al.*, 2019; Smith *et al.*, 2019), and previous work has shown that mutant U2AF1 has an altered binding affinity with its pre-mRNA substrate (Okeyo-Owuor *et al.*, 2015; Fei *et al.*, 2016). In ADC, mutant *U2AF1* has been shown to alter pre-mRNA splicing and other post-transcriptional processes (Brooks *et al.*, 2014; Fei *et al.*, 2016; Park *et al.*, 2016; Palangat *et al.*, 2019).

The impact of *U2AF1* mutations on the transcriptome raises interesting hypotheses for an oncogenic role through mRNA dysregulation. *U2AF1 S34F* is known to alter alternative-splicing and polyadenylation of cancer-relevant genes (Przychodzen *et al.*, 2013; Brooks *et al.*, 2014; Ilagan *et al.*, 2015; Okeyo-Owuor *et al.*, 2015; Shirai *et al.*, 2015; Yip *et al.*, 2015, 2017; Fei *et al.*, 2016; Park *et al.*, 2016; Smith *et al.*, 2019). In myelodysplastic syndromes, *U2AF1 S34F* alters pre-mRNA splicing of Interleukin-1 receptor-associated kinase 4 (IRAK4), producing isoforms that promote activation of kappa-light-chain-enhancer of B cells (NF-kB), a factor known to promote leukemic cell growth (Smith *et al.*, 2019). In addition to splicing-dependent functions of *U2AF1 S34F*, recent studies show a splicing-independent post-transcriptional role for *U2AF1 S34F* in modulating translational efficiency of genes involved in inflammation and metastasis in human bronchial epithelial cells (Palangat *et al.*, 2019). Although some oncogenic roles for *U2AF1 S34F* have been described, the full functional impact of *U2AF1*-associated mRNAs are unknown.

Investigating mRNA isoform function proves difficult given the complexity and accuracy of isoform assembly with short-reads (Engström *et al.*, 2013; Steijger *et al.*, 2013). Accurate isoform assembly is important in investigating RNA processing alterations associated with global splicing factors, like U2AF1. Recent studies have shown the utility of long-read approaches in capturing full-length mRNA isoforms, by constructing isoforms missed by short-read assembly methods (Oikonomopoulos *et al.*, 2016; Byrne *et al.*, 2017; de Jong *et al.*, 2017; Tang *et al.*, 2018; Workman *et al.*, 2019). Moreover, long-read approaches have already been conducted using RNA derived from primary tumor samples harboring *SF3B1* mutations, demonstrating its effectiveness in capturing mutant splicing factor transcriptome alterations *(Tang et al., 2018)*. In addition, studies have shown the extent to which long-read data can be used as a quantitative measure for gene expression (Oikonomopoulos *et al.*, 2016; Byrne *et al.*, 2017). Given the global impact of *U2AF1* mutations on the transcriptome, identifying RNA processing alterations at the level of full-length mRNA isoforms is an essential step in understanding the functional impact of affected mRNAs.

Here we used an emerging long-read sequencing approach to characterize isoform structure and function of transcript isoforms affected by *U2AF1 S34F*. We chose to study *U2AF1 S34F*-associated isoform changes in an isogenic cell line, HBEC3kt cells, which has been used as a model for identifying transcriptome changes associated with *U2AF1 S34F* (Ramirez et. al. 2004, Fei et. al. 2016). We constructed a long-read transcriptome that contains substantial novel mRNA isoforms not reflected in annotations, nor could they be reconstructed using short-read sequencing assembly approaches. Our long-read data supports a strong *U2AF1 S34F* splicing phenotype, in which we demonstrate the ability to recapitulate the splicing phenotype associated with *U2AF1 S34F* mutants using splicing event-level analyses. We find that isoforms containing premature termination codons (PTCs) and immune-related genes are significantly downregulated. Finally, we leverage previously published short-read polysome profiling data to show changes in translational control for genes affected by

U2AF1 S34F. Our work provides additional insights into the function of transcripts altered by *U2AF1 S34F* mutation.

Results

Long-read sequencing reveals the complexity of the HBEC3kt transcriptome

We first characterized the transcriptome complexity of HBEC3kt cells with and without *U2AF1 S34F* mutation using the Oxford Nanopore minION platform. We conducted cDNA sequencing on two clonal cell lines, two wild-type and two *U2AF1 S34F* mutation isolates (WT1, WT2, MT1, MT2). We extracted whole-cell RNA from each cell isolate, one growth replicate of WT1 and M1 and two replicates of WT2 and MT2. We converted RNA into cDNA using methods described in previous nanopore sequencing studies ((Picelli *et al.*, 2013; Byrne *et al.*, 2017)**; Methods**), and performed nanopore 1D cDNA sequencing on individual flow cells (**Figure 1A**). Our sequencing yielded 8.8 million long-reads across all 6 sequencing runs (**Supplemental Table 1**), which we subsequently processed through a Full-Length Alternative Isoform analysis of RNA (FLAIR; (Tang *et al.*, 2018) to construct a reference transcriptome and perform various differential analyses (**Figure 1B, Methods**). We constructed a total of 63,289 isoforms, 49,366 of which were multi-exon and 45,749 contained unique junction sets (**Supplemental Figure 1, Supplemental File 1**).

We compared FLAIR isoforms against GENCODE reference annotations (v19) and a short-read assembly using previously published data from HBEC3kt cells (Fei *et al.*, 2016)**; Supplemental File 2; Methods**). We found that only one-third of our FLAIR transcriptome overlapped with GENCODE annotations (**Figure 2A)**. The remaining FLAIR isoforms contained novel elements, such as novel exons, novel junction combination, or a novel genomic locus. In contrast, nearly half of the isoforms from short-read assembly were comprised of known GENCODE isoforms. We hypothesized that the increased number of annotated isoforms from short-read assembly could be due to higher sequencing depths. We therefore overlapped intron junction-chains between all three datasets and quantified expression from each overlapping group (**Figure 2B**). We found significant differences in expression for isoforms not contained in our set of high confident FLAIR isoforms (p-value <0.001; **Figure 2B** top panel). Although our long-read approach did not capture lowly expressed isoforms, we found a large proportion of FLAIR-exclusive isoforms that contained novel exons, junction combinations and novel loci isoforms (**Figure 2C**). Notably, we identified 182 FLAIR-exclusive isoforms from 123 unannotated loci, none of which were assembled by short-reads despite having short-read coverage support, perhaps due to repeat elements that are known to be difficult to assemble across (Treangen & Salzberg 2012). We investigated a putative lncRNA we call *USFM* (upregulated in splicing factor mutant; *LINC02879*), which was one of the most highly expressed multi-exon isoforms in mutant samples (**Figure 2C,** bottom panel). We manually examined long-reads aligned to *USFM* and found poly(A) tails, suggesting *USFM* supporting reads are not likely to be 3' end fragmented products. Next, we used publicly available ENCODE data to look for transcription factor binding sites and histone modification marks (ENCODE Project Consortium, 2012) that would provide additional evidence of a transcribed gene locus. Peaks associated with H3K27 acetylation and H3K4 methylation suggest the presence of regulated transcribed genomic regions. Moreover, the transcript start site of *USFM* overlapped with active promoter predictions from chromHMM (ENCODE Project Consortium, 2012), an algorithm used to predict promoters and transcriptionally active regions (**Figure 2C,** bottom panel). No significant homology matches to protein-coding domains could be found using NCBI BLAST (Johnson *et al.*, 2008). Taken together, these data indicate that *USFM* isoforms have characteristics that are consistent with lncRNAs, and highlight the utility of long-reads in identifying putative novel genes.

U2AF1 S34F splicing signature captured by long-read event-level analyses

We compared *U2AF1 S34F*-associated splicing signatures in our long-read data to those found from analyses of short-read datasets (Brooks *et al.*, 2014; Ilagan *et al.*, 2015; Okeyo-Owuor *et al.*, 2015; Fei *et al.*, 2016). Previous reports have shown that cassette exon skipping is the most prevalent splicing alteration induced by *U2AF1 S34F*. Moreover, motif analysis of 3' splice sites adjacent to altered cassette exons with enhanced and reduced inclusion show a strong nucleotide context for 'CAG' and 'TAG' acceptor sites, respectively (Brooks *et al.*, 2014; Ilagan *et al.*, 2015; Okeyo-Owuor *et al.*, 2015; Fei *et al.*, 2016). As shown previously, we find that cassette exon events are the most predominant patterns of altered splicing associated with *U2AF1 S34F* mutations in short-read sequencing from The Cancer Genome Atlas (TCGA) lung ADC samples and HBEC3kt isogenic cell lines (179/226 and 187/225, respectively; (Fei *et al.*, 2016). We observe minor differences in alternative donor and intron retention events between TCGA and HBEC3kt short-read, which could likely be explained by the limitations of statistical testing with only two replicates for the HBEC3kt short-read data (Hansen *et al.*, 2011).

We next assayed for alternative-splicing alterations in our long-read data using FLAIR-diffSplice (**Figure 3A**). We used FLAIR for our event-level long-read analysis since most alternative-splicing algorithms were designed for short-read analysis (**Supplemental Table 4**; **Methods**). Our long-read event-level analysis was consistent with short-read analysis. (**Figure 3A**). The most predominant altered splicing pattern from long-read data was cassette exons, in which we found a general trend toward exon exclusion (51/55; **Supplemental Figure 2A**). In addition, we found a good correlation in the magnitude of change in percent spliced in (PSI) between short and long-read PSI values (**Figure 3B;** Pearson r=0.88). Last, we investigated the 3' splice site motif associated with altered cassette exons and alternative acceptor events and found ‘TAG’ and ‘CAG’ motifs associated with acceptor sites with reduced and enhanced inclusion, respectively (**Figure 3C**). Overall, our results demonstrate consistent splicing signature patterns between short and long-read methodologies.

U2AF1 S34F has been implicated in widespread altered poly(A) site selection (Park et. al. 2016). We took advantage of the reduced ambiguity long-reads provide in identifying poly(A) cleavage sites and identified alternative poly(A) alterations associated with *U2AF1 S34F* (**Figure 3D**). Identifying poly(A) sites with short-reads is computationally difficult due to alignment of reads primarily composed of poly(A) sequence, or alignment across repetitive sequence commonly found in 3’ untranslated regions (Chen, Ara and Gautheret, 2009; Elkon, Ugalde and Agami, 2013; Shenker *et al.*, 2015; Ha, Blencowe and Morris, 2018). We first investigated the presence of poly(A) cleavage site motifs at the 3’ ends of FLAIR isoforms, and found a strong signal ~20 nucleotides upstream from transcript end sites for the most commonly used cleavage motif, “AATAAA”, relative to random 6-mer (**Supplemental Figure 2B**). We next tested for APA site usage alterations by comparing the proportion of polyA site usage for each gene between *U2AF1* wild-type and S34F (**Methods**). 10 genes demonstrated significant changes in polyadenylation site usage (corrected p-value <0.05 & ΔAPA > 10%), which comprises 7.2% of all RNA processing alterations identified in this study(11 APA & 142 alternative-splicing events), far less than previous reports. The most significant APA alteration occurred in *BUB3* (**Figure 3E)**, which is part of the mitotic checkpoint pathway, a pathway containing genes that are commonly altered in select lung cancers (Takahashi *et al.*, 1999; Haruki *et al.*, 2001). Collectively, our event-level analyses confirmed our ability to capture well-documented *U2AF1 S34F*-associated splicing signatures with long-read data.

Long-reads provide isoform context for *UPP1* splicing alterations missed by short-read assembly

We compared the exon connectivity of cassette exons altered by *U2AF1 S34F* in Uridine phosphorylase 1 (*UPP1*), which was the most significantly altered gene in our event-level analysis. *UPP1* altered cassette exons accounted for 4 of the 55 significantly altered cassette exons (exons 5, 6-long, 6-short, and 7), one of which, exon 7, was also found to be significantly altered in TCGA ADC data (**Supplemental Table 3**). We compared 28 FLAIR isoforms containing exon 7 against StringTie assembly to determine which isoforms were missed by either method. Despite minor differences in transcript start and end sites, we found all 4 short-read assembled *UPP1* isoforms containing exon 7 in our set of FLAIR isoforms (**Supplemental Figure 2C**). The additional 21 FLAIR-exclusive isoforms contained a mixture of exon skipping events, alternative 3' and alternative 5' splicing events that coincided with exon 7 inclusion. A more broad comparison of all 95 *UPP1* FLAIR isoforms revealed that only 7 were assembled by short-read. We then asked if any of the FLAIR-exclusive *UPP1* isoforms were expressed at substantial proportions (>5%) by quantifying the expression of each isoform using our long-reads. We found that 6 of the 7 most highly expressed *UPP1* isoforms were FLAIR-exclusive (**Supplemental Figure 2D**). Taken together, although short-read methods assembled complex splicing regulation observed in *UPP1,* our long-read analysis revealed extensive isoform diversity not captured by short-reads.

U2AF1 *S34F* induces strong isoform switching in *UPP1* and *BUB3*

We next assessed transcriptome-wide changes in isoform usage using our long-read data. Short-read event-level analyses typically represent isoforms by distinct RNA processing events such as splicing (**Figure 4A**). In contrast, long-reads capture entire mRNA isoforms and therefore can be used to more accurately quantify distinct isoforms. We identified 166 isoforms with significant changes in usage (corrected p-value <0.05) using DRIMSeq (**Supplemental Table 5**, **Methods**). We found that nearly half of significantly altered isoforms (82/166) had large changes in magnitude (Δisoform usage > 10%) (**Figure 4B**). Consistent with our event-level analysis, we found isoforms from *BUB3* and *UPP1* in the top 10 most significantly altered genes, suggesting that changes in these isoforms are defined by splicing event changes. Gene set enrichment analysis using the molecular signature database (Liberzon *et al.*, 2011) on differentially used isoforms (FDR<0.05, Δisoform usage > 10%) revealed genes involved in RNA metabolism, and RNA processing **(Supplemental Table 5**).

We found complex 3' end processing patterns that define *BUB3* isoforms. Previous reports have described alternative acceptor site usage for *BUB3* that leads to the usage of distinct polyadenylation sites (Bava *et al.*, 2013). Consistent with previous reports, we find that the proximal acceptor site leads to the production of isoforms using three APA sites (APA1, 2 and 3), and usage of a distal acceptor site leading to the usage of two APA sites (APA 4, 5) (**Supplemental Figure 3A**). Our event-level analyses revealed a significant shift toward usage of the distal acceptor site and APA site (Δalt. acceptor >30%; corrected p-value < 0.05), which is consistent with differences in isoform usage (**Supplemental Figure 3B**). Notably, the proximal *BUB3* 3' acceptor site is preceded by a thymidine residue, which could partially explain isoform shifting toward the usage of the distal acceptor site. We also observed a preference for APA site 2 in wild-type samples (ΔAPA usage 20%), which seems to be lost in mutant samples (ΔAPA usage 6%; two-sided t-test p-value <0.05).

We next investigated significant isoform usage changes in *UPP1* (**Figure 4C**). Out of the 95 *UPP1* isoforms identified by our data, 68 (71%) fell below 1% of the total *UPP1* gene abundance, indicating that the majority of isoforms are minor isoforms. The remaining 28 *UPP1* isoforms were tested for differential isoform usage, 2 of which were found to have significant usage changes (corrected p-value <0.05 & Δisoform usage >10%) (**Figure 4D showing top 5 expressed isoforms**). RT-PCR validation using primers that span all *U2AF1 S34F*-associated cassette exons showed a pattern consistent with sequencing results, in which *U2AF1 S34F* induces a shift toward *UPP1* isoforms that either contain or exclude all cassette exons (**Figure 4E**). *UPP1* is known to be highly expressed in solid tumors (Lie et. al. 1998; Kanzaki et. al. 2002), but cancer associated splicing alterations have not been described. Altogether, results from our full-length isoform analysis echo results from our event-level analysis, but differentiate which isoforms with similar splicing patterns are actually affected by *U2AF1 S34F.*

<u>Isoforms changes are partially explained by event-level splicing changes</u>

We next determined the extent to which *U2AF1 S34F* alters the expression of individual isoforms. This analysis complements our isoform switching analysis by allowing for the identification of minor isoforms (isoform usage <10%) with large expression changes, and genes with uniform isoform expression changes. Our analysis yielded 122 isoforms with significant changes in expression (corrected p-value<0.05 and log2FoldChange>1.5; **Figure 5A, Supplemental Table 6**). We found the most upregulated isoforms were from the putative lncRNA *USFM* (log2FoldChange > 3 & corrected p-value <0.01). We searched TCGA ADC short-read RNA-seq data for expression of *USFM*, but we could not find substantial read counts (<1 read per million) for samples with or without *U2AF1 S34F* mutations.

In contrast to gene sets identified from our differential isoform usage analysis, we found that differentially expressed isoforms belonged to genes regulated by NF-kB via TNF signaling, most of which are downregulated (corrected p-value <0.05; **Supplemental Table 6**). This observation is particularly important given recent reports implicating *U2AF1 S34F* in altering immune-related genes (Palangat *et al.*, 2019; Smith *et al.*, 2019). We further examined FLAIR isoforms derived from genes in the NF-kB pathway to determine if any *U2AF1 S34F*-associated splicing alterations could explain expression changes. However, when we overlapped results from our event-level splicing analyses along with our gene and isoform expression analyses we found no overlap between NF-kB affected isoforms and *U2AF1 S34F* altered splicing (**Figure 5B**). This result suggests that the expression of these isoforms may be modulated through a splicing-independent mechanism or the altered splicing event cannot be detected by our long-read data.

We expanded our alternative-splicing overlap analysis to ask which of the 198 isoforms with altered usage and expression coincided with other significantly altered features, such as alternative-splicing and gene expression. We found several clusters of features that partially explain the involvement of *U2AF1 S34F* mutation in isoform expression and usage dysregulation. For example, we found 27 isoforms with both significant isoform usage and cassette exon usage changes (**Figure 5C cluster C1**). It is possible that these 27 isoforms with significant usage changes are defined by single exon skipping events and are likely directly induced by *U2AF1 S34F*. In contrast, we found several isoforms that did not overlap any other altered features (**Figure 5C clusters C2 & C7**). Similar to the isoforms from our NF-kb analysis, we suspect these isoform changes to be modulated through either a splicing-independent mechanism or splicing changes undetected by long-reads. We found a single gene, *UPP1*, that contained 4 overlapping features, which were changes in isoform expression, gene expression, cassette exon usage, and isoform usage. Altogether, we observe a consistent pattern of *UPP1* alterations associated with *U2AF1 S34F*, and also identify two populations of dysregulated isoforms that may be modulated through splicing-dependent and independent pathways.

PTC-containing isoforms are downregulated by *U2AF1 S34F*

Our long-read approach enables a more confident open reading frame (ORF) prediction, which can be used to identify altered splicing events that trigger nonsense-mediated decay (NMD). NMD is a process that removes erroneously spliced mRNAs with truncated ORFs that could give rise to gain-of-function or dominant-negative protein products (Dreyfuss, Kim and Kataoka, 2002; Lewis, Green and Brenner, 2003;

Sterne-Weiler *et al.*, 2013; Maslon *et al.*, 2014; Floor and Doudna, 2016; Aviner *et al.*, 2017), and *U2AF1 S34F*-associated spliced products have been shown to be substrates of NMD (Yip *et al.*, 2017). Given the implications of U2AF1 S34F dysregulation and NMD, we asked transcriptome-wide what fraction of altered isoforms could be putative NMD targets. To do this, we classified FLAIR isoforms into two categories, either as putative protein-coding (PRO) isoforms or PTC-containing isoforms (**Methods**; **Figure 6A**). We postulated that the shallow sequencing depth of long-reads relative to short-reads would limit our ability in capturing PTC-containing isoforms if they are indeed subject to NMD. However, of our 63,289 FLAIR isoforms, we identified 8,037 PTC-containing isoforms (12% of all isoforms). We then asked what proportion of PTC-containing isoforms are dysregulated at the level of expression and isoform usage (**Figure 6B**). For differentially used isoforms, we found similar proportions of PTC-containing isoforms (Fisher's exact two-sided test, p=0.5). In contrast, we found a significant difference in the proportion of PTC-containing isoforms between differentially expressed isoforms (Fisher's exact two-sided test, p<0.01). Previous studies have reported S34F-associated splicing alterations that lead to the production of PTC-containing isoforms, yet no study has reported an overall downregulation of such isoforms.

We next sought to test, more generally, if *U2AF1 S34F* induces shifts in isoform productivity by conducting a gene-level analysis. To do this, we compared the proportion of PTC-containing versus productive isoform usage for each gene using the same methodology as our differential isoform usage analysis. Our results showed very few genes with strong shifts in productivity (**Supplemental Figure 4**; 10 total with corrected p-value <0.05 and ΔPTC isoform usage > 10%). However, we did identify a very strong shift in productivity in *UPP1 (*p-value<0.001 & Δproductivity > 20%)*,* a gene we found to have strong changes in splicing, isoform usage, and expression. Interestingly, our differential gene expression analysis showed significant downregulation for *UPP1* (log2FoldChange 2.4 & p-value < 0.05), yet our productivity analysis showed a strong shift toward productive isoform usage (**Figure 6C**). Overall, our results suggest a bias toward downregulation of PTC-containing isoforms in *U2AF1 S34F* cells.

U2AF1 S34F isoform dysregulation is associated with changes in translation

We predicted that if PTC-containing *UPP1* isoform are indeed subject to NMD, then the proportion of *UPP1* mRNAs able to undergo translation will be larger in mutant cells relative to wild-type, since there is a shift toward upregulation of productive *UPP1* isoforms. To test this, we use polysome profiling data from HBEC3kt cells with and without *U2AF1 S34F* causing mutations (**Methods; Figure 6D**). We found a significant change in the proportion of *UPP1* expression across different polysome fractions (chi-squared p-value < 0.01; **Figure 6E**). We observed a large drop (Δ10%) in the proportion of expression in polysome fractions 5 & 6 between mutant and wild-type. These fractions correspond to the monosome, which is a fraction not associated with active translation, and is known to harbor non-coding mRNAs, such as NMD products (Floor and Doudna, 2016). The marked shift of *UPP1* expression in mutant samples from the monosome toward higher polysome (fractions >=7) is consistent with the hypothesis that *U2AF1 S34F*-associated *UPP1* alterations alter mRNA fate by shifting isoform production toward isoforms associated with enhanced translational activity.

We next tested if *U2AF1 S34F*-associated isoform changes in *BUB3* are consistent with differences in polysome profiles. In contrast to *UPP1*, we did not observe significant isoform productivity changes for *BUB3*. Instead, we observed significant changes in a terminal alternative 3' splice site event that is linked to alternative polyadenylation site usage. Previous reports show that *BUB3* APA site 5 is associated with enhanced translational efficiency (Bava *et al.*, 2013). Our APA analysis showed mutant-specific isoform shifts toward isoforms with APA site 5, effectively increasing the proportion of translationally efficient *BUB3* isoforms. We tested for changes in *BUB3* polysome profiles using the same methodology used for *UPP1*. We find a strong shift in *BUB3* expression toward high polysome fractions (**Figure 6F**; chi-squared p-value <0.01). Notably, RNA-IP results from previous reports do not support large changes in cytosolic U2AF1 binding for *BUB3* or *UPP1*, which is a proposed mechanism of mutant *U2AF1* to modulate translational efficiency (Palangat *et al.*, 2019). Altogether, our data indicate a role for translational control through a splicing-dependent manner, and demonstrate distinct mechanisms of *U2AF1 S34F* for modulating translation control of genes through spliced isoform dysregulation.

We next determined if changes in translational control is a general feature for genes with strong changes in isoform expression and usage. Our results showed that 66% (42/63) of genes with *U2AF1 S34F*-associated isoform changes also had a significant change in polysome profile (**Methods**). This proportion was significantly higher (fishers two-sided test p-value <0.01) than the 48% (1340/2753) of genes without S34F-associated isoform changes. Altogether, our results are consistent with previous work implicating *U2AF1 S34F* as a modulator of the translational landscape (Palangat *et al.*, 2019).

Discussion

In this study, we assessed the impact of *U2AF1 S34F*-associated RNA processing alterations on individual mRNAs using an isogenic cell line harboring a *U2AF1 S34F* mutant allele. Although splicing alterations associated with *U2AF1* have been characterized with short-read sequencing, the full-length isoform context in which the altered events occur has not been described. We aimed to fill this gap in knowledge by using a long-read sequencing approach and supplemented our analysis with orthogonal short-read RNA-sequencing datasets from the same isogenic cell lines.

We demonstrate the robustness of long-read approaches by recapitulating splicing signatures associated with *U2AF1 S34F* mutations. Although our long-read transcriptome captures a comparable number of isoforms relative to short-read approaches, we still lack sequencing depth to capture the entire catalog of cassette exons associated with *U2AF1 S34F*, such as known cassette exons in *STRAP* or *ASUN* which were previously described to have *U2AF1 S34F*-associated splicing alterations (Fei *et al.*, 2016). Moreover, although we identified genes with significant changes in polyadenylation site selection, we were unable to recapitulate transcriptome-wide levels observed in previous studies (Park et. al. 2016). In line with these shortcomings, a saturation analysis of full-length isoforms construction reveals isoform discovery limitations, possible due to relatively shallow sequencing depth (**Supplemental Figure 6**). However, long-read sequencing approaches offered by PacBio and Oxford nanopore are continually improving sequencing throughput and quality. Recent studies using newer Nanopore flow cell chemistry and higher-throughput platforms have demonstrated data yield orders of magnitude greater than this study (Tang *et al.*, 2018). With greater data yield and improved transcriptome coverage, there is the potential to identify more *U2AF1 S34F* dysregulated isoforms with greater confidence.

We observe an interesting link between isoform dysregulation and translational control. Previous studies using RNA immunoprecipitation assays have shown that cytosolic mRNA binding of U2AF1 can modulate translational control (Palangat *et al.*, 2019). This splicing-independent mechanism of translational control is complementary to our findings here, in which isoforms arising from RNA processing alterations caused by *U2AF1 S34F* cause changes in translational control of the gene. Interestingly, our data supports two potential mechanisms. In the case of *BUB3*, *U2AF1 S34F* induces isoform switches toward isoforms with regulatory sequences that promote high translational efficiency. Alternatively, for *UPP1* we observe a substantial shift away from PTC-containing isoforms, which could serve as putative NMD targets. While further studies are necessary to directly test if these PTC-containing isoforms are regulated by NMD, we hypothesize that the expression of PTC-containing isoforms is strongly selected against in the presence of *U2AF1 S34F*.

Our analyses contribute several findings implicating *UPP1* as severely dysregulated by *U2AF1 S34F*. So far, no reports have mentioned isoform-specific dysregulation associated with *UPP1*. *UPP1* has been observed to be upregulated in certain cancer types (Liu *et al.*, 1998). In our study using non-cancer derived cells, we find an opposite pattern, in which *UPP1* is significantly downregulated at the level of overall gene expression. The observed downregulation of *UPP1* is consistent with our finding of downregulation of isoforms involved in the TNF via NF-kB signaling pathway, which is a positive regulator of *UPP1* expression (Wan *et al.*, 2006). However, although we observe a strong downregulation at the level of total gene expression, our isoform usage and productivity analyses reveal a shift toward more productive isoforms. Nevertheless, further studies are required to determine what impacts *UPP1* isoform changes have on cellular function.

Overall, our data captured the context in which *U2AF1 S34F* RNA processing alterations occur at full-length isoform resolution. We build upon previous short-read analyses by providing an extensive list of isoform-specific changes associated with *U2AF1 S34F*, along with the first estimates of isoform function. Our results demonstrate the importance of investigating the transcriptome of mutant splicing factors using long-read data that provides diverse perspectives on RNA processing and isoform function.

Methods

Data generation and processing

Preparing RNA for long-read sequencing

HBEC3kt cells with and without *U2AF1 S34F* were cultured as previously described (Ramirez *et al.*, 2004; Fei *et al.*, 2016). Total RNA was extracted from whole cell lysate using Zymo Direct-zol RNA kits. Purified RNA was prepared for long-read following previously established protocols (Picelli *et al.*, 2013; Byrne *et al.*, 2017; Tang *et al.*, 2018). Total RNA was reverse transcribed using the SmartSeq2 protocol, and amplified using 15 cycles of PCR. 1 ug of PCR amplified cDNA from each sample was subsequently used for Oxford Nanopore 1D library preparation (SQK-LSK108) on flow cell chemistry version 9.4. Basecalling was performed using Albabacore version 2.1.0 using options --flowcell FLO-MIN106 and --kit SQK-LSK108. Nanopore reads were prepared for genomic alignment by removing adapter sequenced using Porechop version 0.2.3 (Wick, 2017). After adapter removal, reads were aligned to GENCODE hg19 using minimap2 version 2.14-r894-dirty (Li, 2018) using the `-ax` option.

Processing TCGA LUAD short-read data

Lung adenocarcinoma short-read data from The Cancer Genome Atlas (601 samples total) was downloaded from CGhub using gtdownload (Wilks *et al.*, 2014). TCGA donors with multiple RNA-seq bams were filtered by date to only include the most recent RNA-seq bam (495 samples). 495 TCGA bams were subsequently processed through JuncBase using default parameters with GENCODE hg19 comprehensive annotations and basic annotations as input to `getASEventReadCounts` for options `--txt_db1` and `txt_db2`, respectively (Brooks *et al.*, 2011). Differential splicing analyses were performed using Wilcoxon-rank sum between samples containing *U2AF1 S34F* splicing factor mutation (n=11) or no splicing factor mutation (n=451), which were defined by molecular profiling details outlined in Campbell et. al. (Campbell *et al.*, 2016).

Obtaining and processing HBEC3kt short-read data

Short-read HBEC3kt data was retrieved from NCBI short read archive (GSE80136). Reads were aligned to GENCODE hg19 using STAR version 2.5.3a (Dobin *et al.*, 2013) with parameters `--twopassMode Basic`. Aligned bams were subsequently individually used for transcriptome assembly using StringTie version 1.3.5 using GENCODE hg19 basic annotations (Pertea *et al.*, 2015). Individual GTF annotation files generated from StringTie were then merged using default parameters. For the differential splicing analysis of HBEC3kt short-read data, we used JuncBASE with the same methodology as described in TCGA LUAD short-read data methods section. HBEC3kt short-read data had two biological replicates per condition (wild-type and mutant); therefore, for statistical testing, we conducted pairwise fisher's tests, then defined significant events as ones with a Benjamini-Hochberg corrected p-value > 0.05 within each condition and a corrected p-value <0.05 between samples across conditions. We then post-filtered significant events to remove redundant and overlapping events by running JuncBASE scripts `makeNonRedundantAS.py` and `getSimpleAS.py`. To compare long and short-read Δpercent spliced-in values (PSI), we computed PSI changes for significant long-read cassette exon events by subtracting DRIMSeq-calculated proportion values for wild-type and mutant. We then filtered our short-read JuncBASE PSI table for significant long-read events, and computed the short-read change in PSI by subtracting the average PSI between wild-type and mutant.

Long-read Analysis

Nanopore read correction, FLAIR-correct

Aligned Nanopore sequencing data were concatenated prior to running FLAIR v1.4 (Tang *et al.*, 2018) using samtools v 1.9 (Li *et al.*, 2009). Bam files were converted to bed using FLAIR-bam2bed12. Converted bed alignments were subsequently corrected using `FLAIR-correct` with GENCODE hg19 basic annotations. Junctions identified by STAR alignment of HBEC3kt short-read data were also used as input into FLAIR-correct. Briefly, STAR junctions were kept if they contained at least 3 uniquely aligned in either both Mut1a and Mut1b samples or in both WT1 and WT2 samples. Junctions that did not follow GT-AG splicing motif were also removed.

FLAIR-collapse and diffExp

Differential analyses were performed by FLAIR-diffExp with default parameters. Genes and isoforms with less than 10 reads from either sample group were excluded from isoform expression and usage analyses.

Long-read alternative-splicing analysis, FLAIR-diffSplice

Differential alternative splicing for long-read data was conducted with FLAIR-diffSplice. FLAIR-diffSplice calls events for the following alternative-splicing types: cassette exon usage, alternative 3' splice site, alternative 5' splice site, intron retention, and alternative polyadenylation. Percent spliced-in values for each event were calculated by tallying the number of reads supporting isoforms that include an event, divided by the total number of reads that span the event. Inclusion and exclusion counts were then constructed into a table to process with DRIM-seq (Nowicka et. al. 2016) for differential splicing analysis.

Long-read alternative polyadenylation analysis

Poly(A) cleavage sites were defined by clustering FLAIR isoform transcript end sites using BedTools cluster, with a window distance of 5 (Quinlan and Hall, 2010; Quinlan, 2014). Poly(A) sites were then quantified by summing the total number of aligned read counts for each isoform that fell within each cluster. Clusters were assigned to genes, and counts for each cluster were then processed by DRIM-Seq.

Genes with corrected p-value < 0.05 were considered to have significant changes in poly(A) site usage.

Gene-set enrichment analysis

The Molecular Signatures Database (Liberzon *et al.*, 2011, 2015) was used to perform all gene set enrichment analysis using gene sets: GO gene sets, Hallmarks and Canonical pathways. Genes names included for isoform expression and isoform usage analyses were from isoforms with corrected p-value <0.05, and magnitude changes of ΔLog2FoldChange>1.5 and Δ10% isoform usage. Duplicate gene names from genes with multiple significantly altered isoforms were included only once.

Polysome analysis

Polysome profiling data from HBEC3kt cells with and without *U2AF1 S34F* mutation were obtained from Palangat et. al. (2019; Supplemental Table S4). For each gene, normalized read counts across polysome fractions 3 through 10-12 were compared between mutant and wild type samples using Chi-squared test. Genes with less than 11 normalized read counts in any given fraction were not tested. Multiple testing correction was conducted using the python module statsmodels.stats.multitest.multipletests with default parameters. Significant changes in polysome profile were considered to have a corrected p-value of <0.05. We tested for general polysome profile alterations in *U2AF1 S34F*-associated genes by comparing the ratio of affected genes with and without significant changes in polysome profile versus unaffected genes. Affected genes were considered ones with either a significant isoform expression or usage change.

Statistics and significance testing

Results from all differential analyses were called significant if their corrected p-value fell below $p<0.05$ and passed a magnitude filter. For differentially expressed isoforms, events over a log2 fold change of 1.5 were called significant. For differentially used isoforms and alternative splicing events, events with >=10% change in usage were called significant.

FIGURES

FIGURE 1

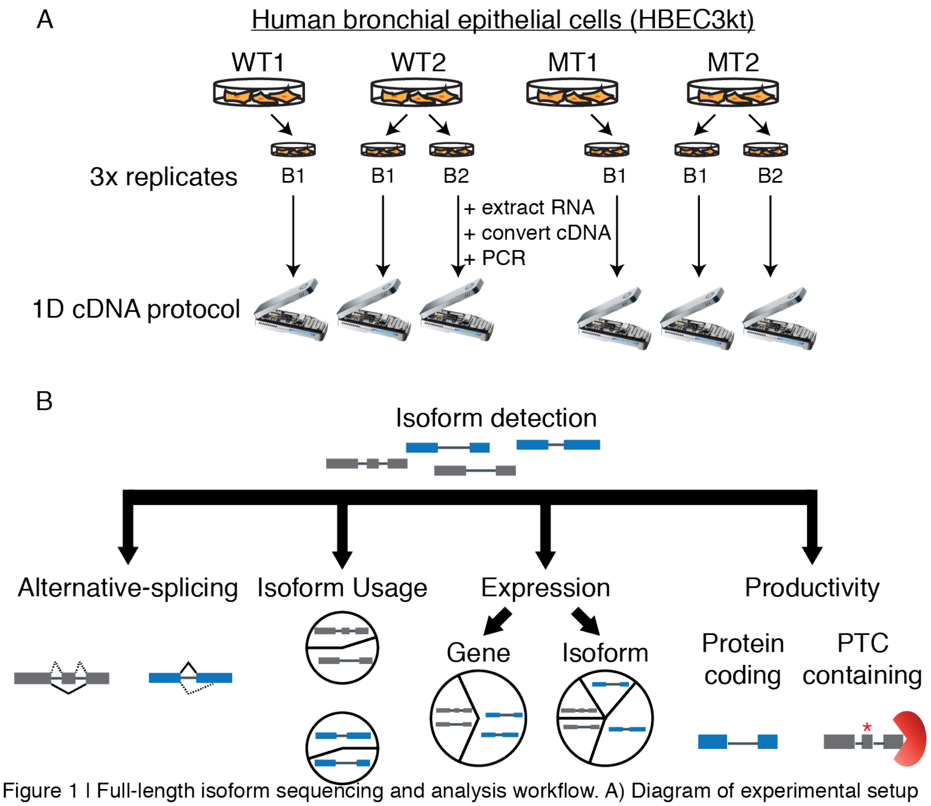

Figure 1 I Full-length isoform sequencing and analysis workflow. A) Diagram of experimental setup and sequencing strategy. RNA was extracted from whole cell lysate and converted to cDNA using a polyA tail selection strategy. Wild-type and mutant conditions were sequenced in triplicate. Each sequencing run was conducted in parallel, in which a wild-type or mutant was sequenced on separate flow cells. B) Data processing pipeline workflow. FLAIR was used to construct a reference transcriptome from long-read data with matched short-read RNA-seq and to perform differential expression and productivity analyses.

FIGURE 2

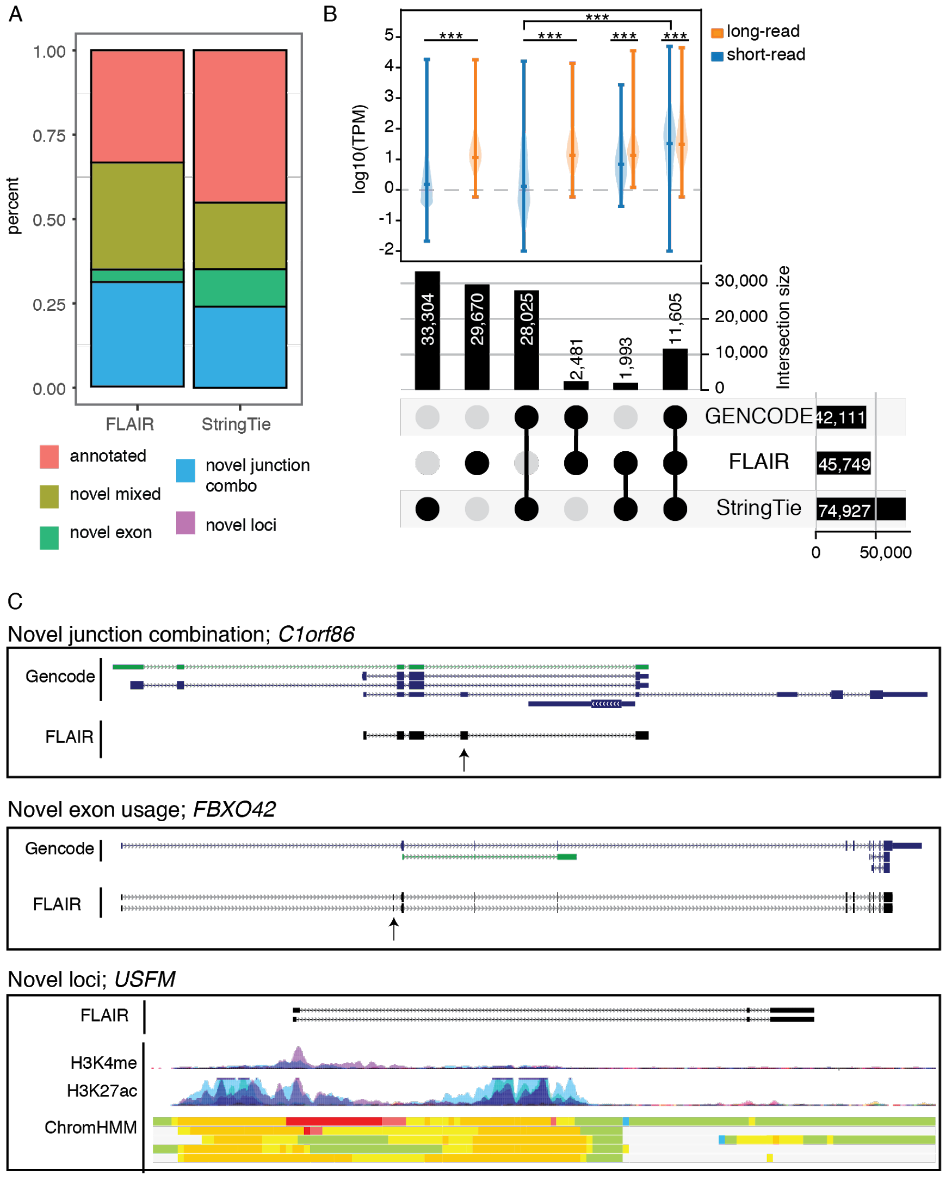

Figure 2 I FLAIR captures HBEC3kt transcriptome complexity A) Isoform annotation categories for FLAIR and StringTie isoforms in comparison to GENCODE v19 annotations. B) Overlap of transcript isoforms between long-read FLAIR, short-read StringTie assembly, and Gencode v19 annotation. Top panel, expression quantification from StringTie and FLAIR-quantify for isoforms in each overlap category. Expression distributions were compared using wilcoxon rank sum test, and comparisons denoted with *** have p-values < 0.001. C) UCSC genome browser shot example for novel classification categories. For each panel, Gencode annotations represent gencode v19 basic annotation set. For Novel Exon and Novel Loci panels, Encode regulatory tracks were included to show H3K27 acetylation, DNase hypersensitivity, Transcription factor binding ChIP (TF Chip), and ChromHMM data from various cell lines. Red and yellow hues represent putative promoter regions; Green regions represent putative transcribed regions.

FIGURE 3

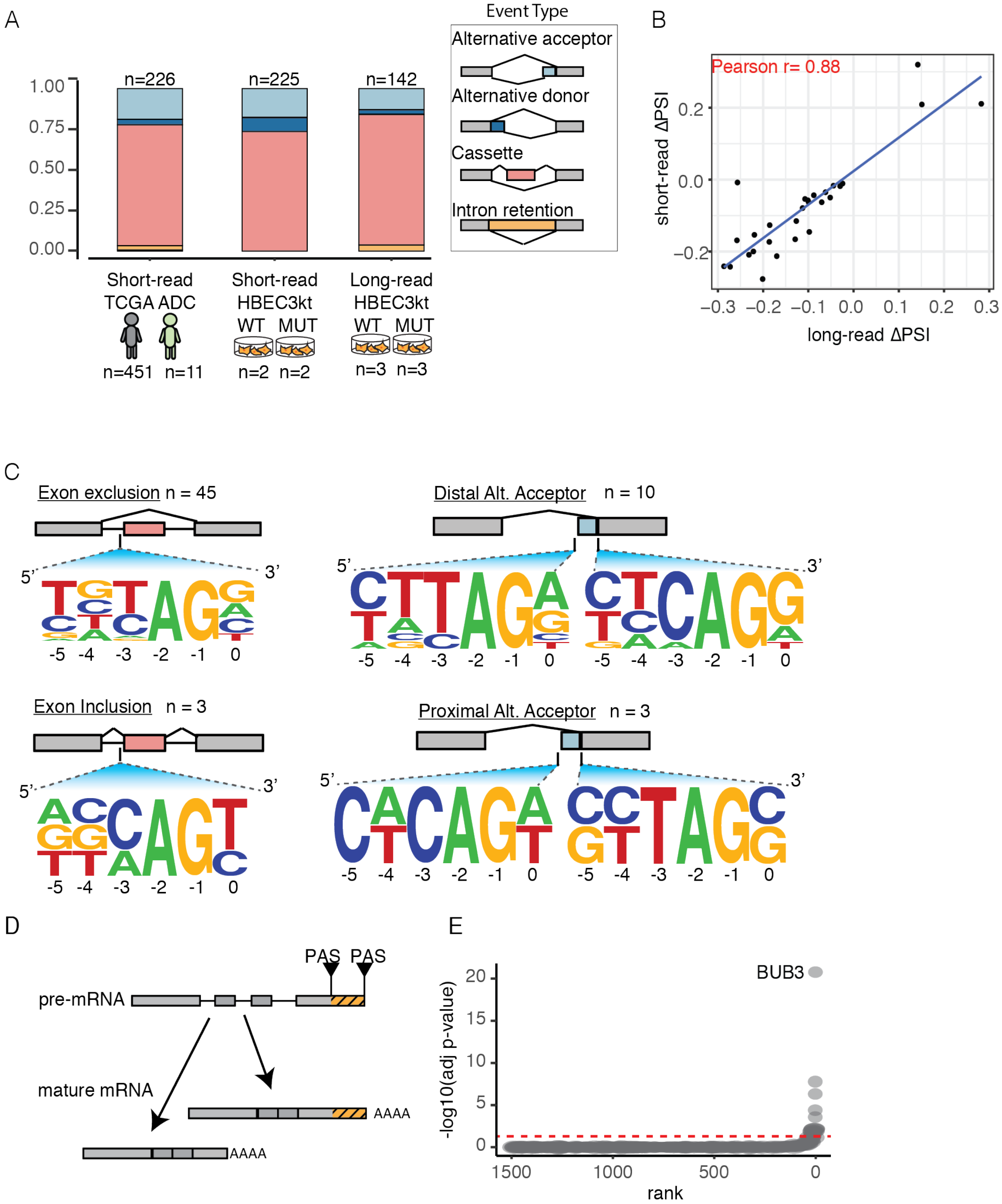

Figure 3 I Nanopore data recapitulates *U2AF1 S34F* splicing signature A) Alternative-splicing events that were found to be significantly altered between wild-type and U2AF1 S34F conditions. Events are broken down into different patterns of alternative-splicing B) Change in PSI (percent spliced-in) correlation between short and long-read cassette exon events C) Weblogos of 3' splice sites for altered cassette exons (left panels) and alternative acceptor sites (right panels) identified using nanopore data. D) Alternative polyadenylation (APA) site selection schematic. E) Ranked genes with significant changes in APA site usage.

FIGURE 4

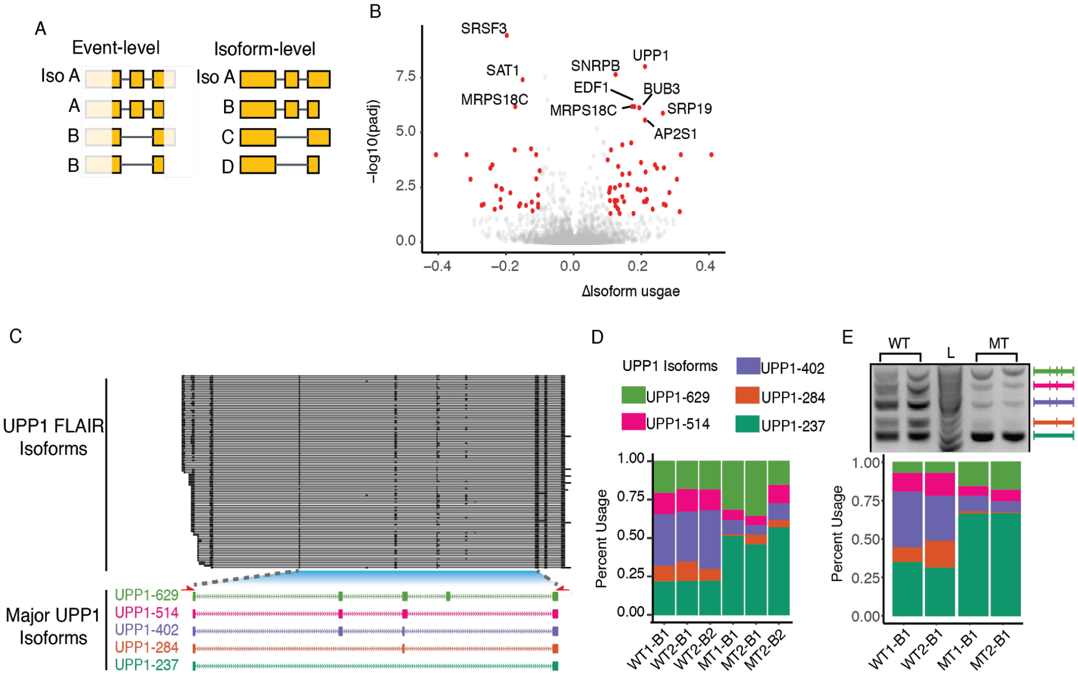

Figure 4 | *U2AF1 S34F*-associated full-length isoform usage changes A) Diagram of event-level versus isoform-level analyses captured by long-read sequencing. alterations. B) Volcano plot of differentially used isoforms. Red dots indicate usage changes with corrected p value<0.05 and magnitude change >10%. Gene names indicate top 10 genes with significant isoform changes. C) UPP1 FLAIR isoforms (top panel), and major isoforms (bottom panel). Isoform numbers correspond to predicted amplion sizes. Red arrows below major isoforms represent PCR primers used for RT-PCR validation. D) Long-read isoform usage quantified by nanopore data. E) 2% agarose gel with UPP1 amplicon products (top panel) and gel quantification bar chart (bottom panel).

FIGURE 5

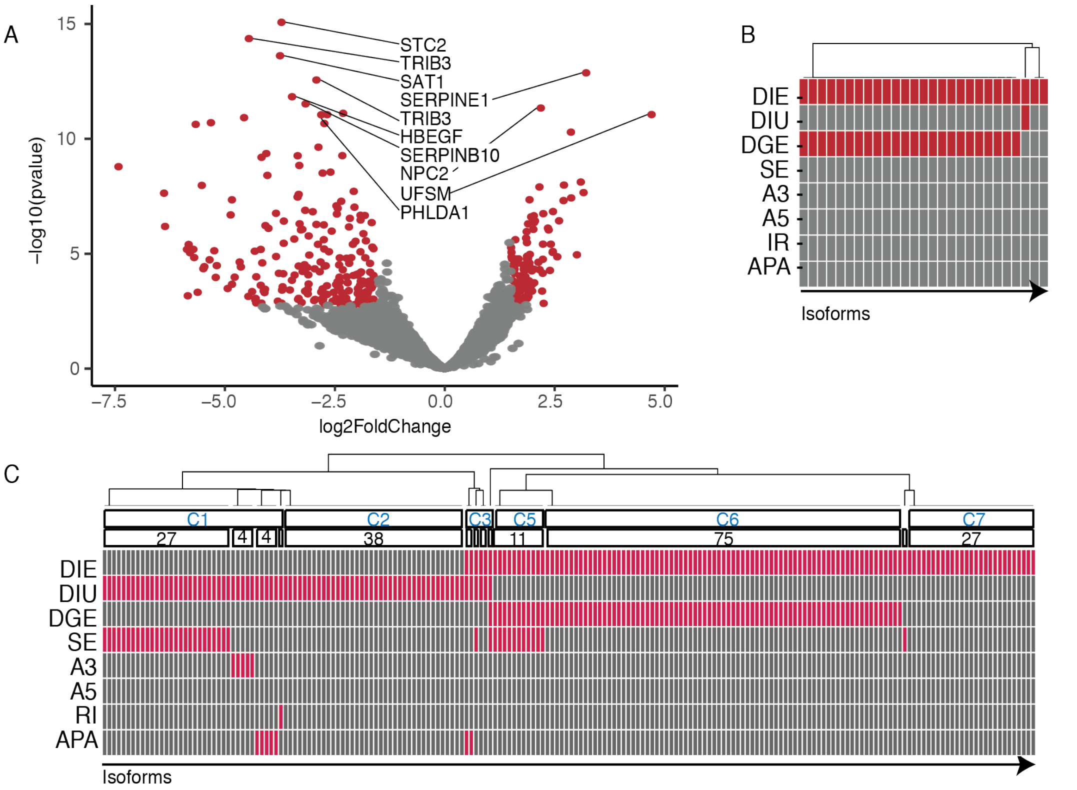

Figure 5 l S34F-associated full-length isoform expression alterations A) Volcano plot of differentially expressed isoforms. Red dots indicate expression changes with adjusted p-value<0.05 and magnitude change >1.5. Gene names are ordered by top 10 most significantly altered isoforms. B) Differential-event overlap of isoforms from genes involved NF-kb signaling pathway. Each box indicates an isoform, where red signifies if a particular isoform resulted as significantly altered for the corresponding analysis. DIE, differential isoform expression; DIU, differential isoform usage; DGE, differential gene expression; SE, skipped exon; A3, alternative 3' splice site usage; A5, alternative 5' splice site usage ; IR, intron retention; APA, alternative polyadenylation site usage. C) Same as panel D, except including all isoforms with altered expression or usage.

FIGURE 6

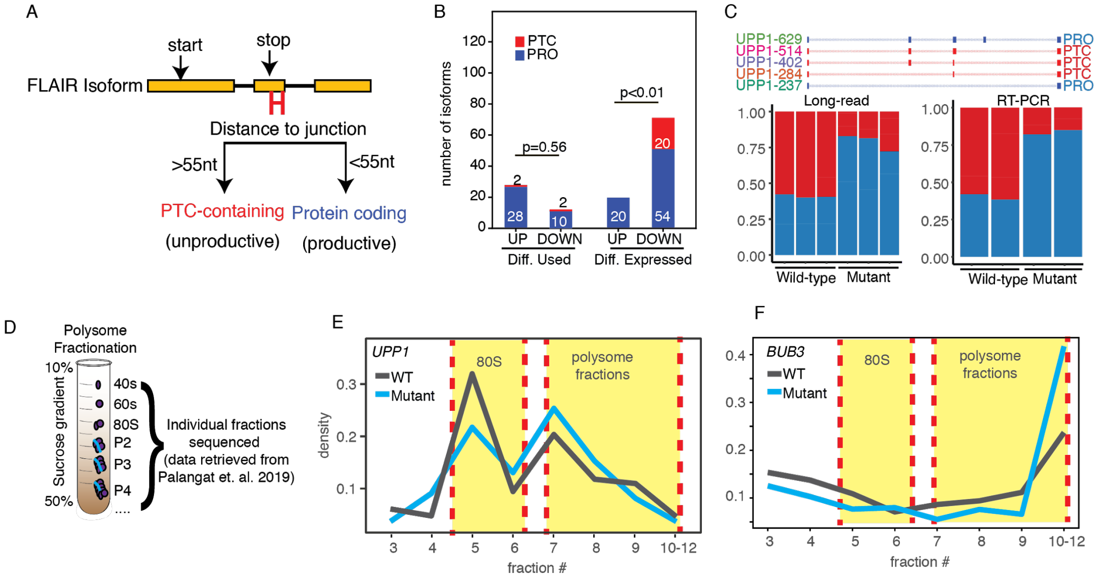

Figure 6 | *U2AF1 S34F* induces shifts in isoform productivity A) Diagram of isoform productivity logic prediction. B) Comparison of up- and down-regulated S34F-associated isoform changes classified by productivity. UP, upregulated; DOWN, downregulated. Upregulated indicate isoforms with increased usage frequency or expression relative to wild-type. Downregulated indicate isoforms with decreased usage frequency or expression relative to wild-type. C) *UPP1* major isoforms classified by productivity. Bar plots show quantification of each productivity type for nanopore long-read data (left panel), and RT-PCR quantification (right panel). D) Polysome profiling analysis and sequencing scheme. E) *UPP1* expression density of normalized read counts across polysome fractions. Yellow highlights indicate 80s and polysome fractions. F) Same as panel E for *BUB3*.

Acknowledgments

We thank Dennis Fei and Harold Varmus for providing the HBEC3kt cells. We also thank the ENCODE Consortium and the ENCODE production lab of Bradley E. Bernstein and the Kellis Computational Biology group for production of the histone regulatory data used via the UCSC Genome browser tracks for this study. This work was supported by the Pew Charitable Trusts and the University of California Tobacco-Related Disease Research Program T29KT040I (A.N.B.). A.D.T. was funded through NIH grant 5T32HG008345. C.M.S. was supported by training grants NIH T32GM008646, 2R25GM058903, and the Ford Foundation predoctoral fellowship.

Author Contribution

C.M.S. and A.N.B. designed the study. C.M.S. and E.H.R designed experiments. C.M.S performed experiments. C.M.S., A.D.T., M.G.M., A.N.B. wrote code and analyzed data. C.M.S., E.H.R., A.D.T., A.N.B. interpreted the data. C.M.S. and A.N.B. wrote the manuscript with input from all other co-authors.

Conflict of Interest

A.N.B. and A.D.T. has been reimbursed for travel, accommodation, and registration for conference sessions organized by Oxford Nanopore Technologies.

Code Availability

All FLAIR related scripts and modules used in this study can be found at https://github.com/BrooksLabUCSC/FLAIR. FLAIR commands and other code are available as jupyter notebooks upon request.

Data Availability

Long-read nanopore sequencing data from HBEC3kt wild type and *U2AF1 S34F* cells are available in the NCBI GEO database (GSE140734 accession number).

References

Aviner, R. *et al.* (2017) 'Proteomic analysis of polyribosomes identifies splicing factors as potential regulators of translation during mitosis', *Nucleic acids research*. academic.oup.com, 45(10), pp. 5945–5957.

Bava, F.-A. *et al.* (2013) 'CPEB1 coordinates alternative 3'-UTR formation with translational regulation', *Nature*. nature.com, 495(7439), pp. 121–125.

Brooks, A. N. *et al.* (2011) 'Conservation of an RNA regulatory map between Drosophila and mammals', *Genome research*. genome.cshlp.org, 21(2), pp. 193–202.

Brooks, A. N. *et al.* (2014) 'A pan-cancer analysis of transcriptome changes associated with somatic mutations in U2AF1 reveals commonly altered splicing events', *PloS one*. journals.plos.org, 9(1), p. e87361.

Byrne, A. *et al.* (2017) 'Nanopore long-read RNAseq reveals widespread transcriptional variation among the surface receptors of individual B cells', *Nature communications*. nature.com, 8, p. 16027.

Campbell, J. D. *et al.* (2016) 'Distinct patterns of somatic genome alterations in lung adenocarcinomas and squamous cell carcinomas', *Nature genetics*. nature.com, 48(6), pp. 607–616.

Cancer Genome Atlas Research Network (2014) ‘Comprehensive molecular profiling of lung adenocarcinoma’, *Nature*. nature.com, 511(7511), pp. 543–550.

Chen, C., Ara, T. and Gautheret, D. (2009) ‘Using Alu elements as polyadenylation sites: A case of retroposon exaptation’, *Molecular biology and evolution*. academic.oup.com, 26(2), pp. 327–334.

Coulon, A. *et al.* (2014) ‘Kinetic competition during the transcription cycle results in stochastic RNA processing’, *eLife*. cdn.elifesciences.org, 3. doi: 10.7554/eLife.03939.

Dobin, A. *et al.* (2013) ‘STAR: ultrafast universal RNA-seq aligner’, *Bioinformatics* . academic.oup.com, 29(1), pp. 15–21.

Dreyfuss, G., Kim, V. N. and Kataoka, N. (2002) ‘Messenger-RNA-binding proteins and the messages they carry’, *Nature reviews. Molecular cell biology*. nature.com, 3(3), pp. 195–205.

Elkon, R., Ugalde, A. P. and Agami, R. (2013) ‘Alternative cleavage and polyadenylation: extent, regulation and function’, *Nature reviews. Genetics*, 14(7), pp. 496–506.

ENCODE Project Consortium (2012) ‘An integrated encyclopedia of DNA elements in the human genome’, *Nature*, 489(7414), pp. 57–74.

Engström, P. G. *et al.* (2013) ‘Systematic evaluation of spliced alignment programs for RNA-seq data’, *Nature methods*, 10(12), pp. 1185–1191.

Fei, D. L. *et al.* (2016) ‘Wild-Type U2AF1 Antagonizes the Splicing Program Characteristic of U2AF1-Mutant Tumors and Is Required for Cell Survival’, *PLoS genetics*. journals.plos.org, 12(10), p. e1006384.

Floor, S. N. and Doudna, J. A. (2016) ‘Tunable protein synthesis by transcript isoforms in human cells’, *eLife*. cdn.elifesciences.org, 5. doi: 10.7554/eLife.10921.

Ha, K. C. H., Blencowe, B. J. and Morris, Q. (2018) ‘QAPA: a new method for the systematic analysis of alternative polyadenylation from RNA-seq data’, *Genome biology*, 19(1), p. 45.

Hansen, K. D. *et al.* (2011) ‘Sequencing technology does not eliminate biological variability’, *Nature biotechnology*, 29(7), pp. 572–573.

Haruki, N. *et al.* (2001) ‘Molecular analysis of the mitotic checkpoint genes BUB1, BUBR1 and BUB3 in human lung cancers’, *Cancer letters*, 162(2), pp. 201–205.

Ilagan, J. O. *et al.* (2015) ‘U2AF1 mutations alter splice site recognition in hematological malignancies’, *Genome research*. genome.cshlp.org, 25(1), pp. 14–26.

Johnson, M. *et al.* (2008) ‘NCBI BLAST: a better web interface’, *Nucleic acids research*. academic.oup.com, 36(Web Server issue), pp. W5–9.

de Jong, L. C. *et al.* (2017) ‘Nanopore sequencing of full-length BRCA1 mRNA transcripts reveals co-occurrence of known exon skipping events’, *Breast cancer research: BCR*, 19(1), p. 127.

Lewis, B. P., Green, R. E. and Brenner, S. E. (2003) ‘Evidence for the widespread coupling of alternative splicing and nonsense-mediated mRNA decay in humans’, *Proceedings of the National Academy of Sciences of the United States of America*. National Acad Sciences, 100(1), pp. 189–192.

Liberzon, A. *et al.* (2011) ‘Molecular signatures database (MSigDB) 3.0’, *Bioinformatics* . academic.oup.com, 27(12), pp. 1739–1740.

Liberzon, A. *et al.* (2015) ‘The Molecular Signatures Database (MSigDB) hallmark gene set collection’, *Cell systems*, 1(6), pp. 417–425.

Li, H. *et al.* (2009) ‘The Sequence Alignment/Map format and SAMtools’, *Bioinformatics* . academic.oup.com, 25(16), pp. 2078–2079.

Li, H. (2018) ‘Minimap2: pairwise alignment for nucleotide sequences’, *Bioinformatics* . academic.oup.com, 34(18), pp. 3094–3100.

Liu, M. *et al.* (1998) ‘Expression, characterization, and detection of human uridine phosphorylase and identification of variant uridine phosphorolytic activity in selected human tumors’, *Cancer research*. AACR, 58(23), pp. 5418–5424.

Maslon, M. M. *et al.* (2014) ‘The translational landscape of the splicing factor SRSF1 and its role in mitosis’, *eLife*. cdn.elifesciences.org, p. e02028.

Oikonomopoulos, S. *et al.* (2016) ‘Benchmarking of the Oxford Nanopore MinION sequencing for quantitative and qualitative assessment of cDNA populations’, *Scientific reports*. nature.com, 6, p. 31602.

Okeyo-Owuor, T. *et al.* (2015) ‘U2AF1 mutations alter sequence specificity of pre-mRNA binding and splicing’, *Leukemia*. nature.com, 29(4), pp. 909–917.

Palangat, M. *et al.* (2019) ‘The splicing factor U2AF1 contributes to cancer progression through a noncanonical role in translation regulation’, *Genes & development*. genesdev.cshlp.org, 33(9-10), pp. 482–497.

Park, S. M. *et al.* (2016) ‘U2AF35(S34F) Promotes Transformation by Directing Aberrant ATG7 Pre-mRNA 3’ End Formation’, *Molecular cell*. Elsevier, 62(4), pp. 479–490.

Pertea, M. *et al.* (2015) ‘StringTie enables improved reconstruction of a transcriptome from RNA-seq reads’, *Nature biotechnology*. nature.com, 33(3), pp. 290–295.

Picelli, S. *et al.* (2013) ‘Smart-seq2 for sensitive full-length transcriptome profiling in single cells’, *Nature methods*. nature.com, 10(11), pp. 1096–1098.

Przychodzen, B. *et al.* (2013) ‘Patterns of missplicing due to somatic U2AF1 mutations in myeloid neoplasms’, *Blood*. Am Soc Hematology, 122(6), pp. 999–1006.

Quinlan, A. R. (2014) ‘BEDTools: the Swiss-army tool for genome feature analysis’, *Current protocols in bioinformatics / editoral board, Andreas D. Baxevanis ... [et al.]*. Wiley Online Library, 47(1), pp. 11–12.

Quinlan, A. R. and Hall, I. M. (2010) ‘BEDTools: a flexible suite of utilities for comparing genomic features’, *Bioinformatics* . academic.oup.com, 26(6), pp. 841–842.

Ramirez, R. D. *et al.* (2004) ‘Immortalization of human bronchial epithelial cells in the absence of viral oncoproteins’, *Cancer research*. AACR, 64(24), pp. 9027–9034.

Shao, C. *et al.* (2014) ‘Mechanisms for U2AF to define 3’ splice sites and regulate alternative splicing in the human genome’, *Nature structural & molecular biology*. nature.com, 21(11), pp. 997–1005.

Shenker, S. *et al.* (2015) ‘IsoSCM: improved and alternative 3’ UTR annotation using multiple change-point inference’, *RNA* , 21(1), pp. 14–27.

Shirai, C. L. *et al.* (2015) ‘Mutant U2AF1 Expression Alters Hematopoiesis and Pre-mRNA Splicing In Vivo’, *Cancer cell*. Elsevier, 27(5), pp. 631–643.

Smith, M. A. *et al.* (2019) ‘U2AF1 mutations induce oncogenic IRAK4 isoforms and activate innate immune pathways in myeloid malignancies’, *Nature cell biology*. nature.com, 21(5), pp. 640–650.

Steijger, T. *et al.* (2013) ‘Assessment of transcript reconstruction methods for RNA-seq’, *Nature methods*. nature.com, 10(12), pp. 1177–1184.

Sterne-Weiler, T. *et al.* (2013) ‘Frac-seq reveals isoform-specific recruitment to polyribosomes’, *Genome research*. genome.cshlp.org, 23(10), pp. 1615–1623.

Takahashi, T. *et al.* (1999) ‘Identification of frequent impairment of the mitotic checkpoint and molecular analysis of the mitotic checkpoint genes, hsMAD2 and p55CDC, in human lung cancers’, *Oncogene*. nature.com, 18(30), pp. 4295–4300.

Tang, A. D. *et al.* (2018) ‘Full-length transcript characterization of SF3B1 mutation in chronic lymphocytic leukemia reveals downregulation of retained introns', *bioRxiv*. doi: 10.1101/410183.

Wan, L. *et al.* (2006) 'Modulation of uridine phosphorylase gene expression by tumor necrosis factor-α enhances the antiproliferative activity of the capecitabine intermediate 5′-deoxy-5-fluorouridine in breast cancer cells', *Molecular pharmacology*. ASPET, 69(4), pp. 1389–1395.

Wick, R. (2017) 'Porechop'. Github https://github. com/rrwick/Porechop.

Wilks, C. *et al.* (2014) 'The Cancer Genomics Hub (CGHub): overcoming cancer through the power of torrential data', *Database: the journal of biological databases and curation*. academic.oup.com, 2014. doi: 10.1093/database/bau093.

Workman, R. E. *et al.* (2019) 'Nanopore native RNA sequencing of a human poly(A) transcriptome', *Nature methods*. doi: 10.1038/s41592-019-0617-2.

Yip, B. H. *et al.* (2015) 'Lineage-Specific Aberrant mRNA Splicing By U2AF1 Mutation Alters Erythroid and Granulomonocytic Differentiation', *Blood*. American Society of Hematology, 126(23), pp. 142–142.

Yip, B. H. *et al.* (2017) 'The U2AF1S34F mutation induces lineage-specific splicing alterations in myelodysplastic syndromes', *The Journal of clinical investigation*. Am Soc Clin Investig, 127(9), p. 3557.

Supplemental Material

Supplemental Table 1 - Nanopore read length and GC statistics. Tab delimited matrix containing statistics for each sequencing run for WT1, WT2, MT1, MT2 samples (biological replicate 1 and 2 - B1 & B2). Columns describe the following: numReads - total number of reads, numBases - total number of bases called for each sequencing run, meanLen - read length statistical average , medLen - read length statistical median, minLen - minimum read length, maxLen - longest read length, meanGC - GC content statistical average, medGC - GC content statistical median.

Supplemental Table 2 - JuncBASE table of TCGA-associated U2AF1 splicing events. Tab delimited matrix containing alternative-splicing events and quantifications from short-read TCGA lung adenocarcinoma RNA-seq data identified by juncBASE. Each entry represents an individual splicing event. Each column represents distinct characteristics for each event: novel_event - describes whether the inclusion coordinate overlaps GENCODE v19 annotated intron boundaries, as_type - describes the alternative-splicing pattern type, hugo - denotes the hugo symbol gene name for each event, chromosome & strand - describe the chromosome and reference event strand, exclusion_coords & inclusion_exon_coord - define the genomic ranges which exclude and include each event, ΔPSI - is the difference in PSI between U2AF1 S34F mutant and non-mutant samples, p-value & p.adj - are the raw and Benjamini Hochberg adjusted p-values from wilcoxon rank sum test.

Supplemental Table 3 - JuncBASE table of short-read HBEC3kt analysis. Tab delimited matrix containing alternative-splicing events and quantifications from short-read HBEC3kt samples from Fei et. al. (2016). Samples are denoted by SRA sample numbers. Columns describe event-type, inclusion and exclusion coordinates (similar to Supplemental Table 2). Values in columns headed by samples SRR numbers denote percent spliced in (PSI) values range from 0 to 100. Events for which a sample did not have at least 25 supporting reads are denoted as “NA”

Supplemental Table 4 - List of significant FLAIR-diffSplice of splicing events. Tab delimited file of alternative splicing events identified in long-read data. Alternative-splicing type and coordinate of event are described in the first column: es - cassette exon, a3 - alternative acceptor, a5 - alternative donor, ir - retained intron. Remaining columns describe the magnitude change in percent spliced in (PSI) as determined by DRIM-seq and significance of change.

Supplemental Table 5 - List of differentially used FLAIR isoforms and enriched gene sets.
Table 1 corresponds to differentially used isoforms as determined by DESeq2. Log2FoldChanges represent shrinkage computed changes as computed by the LFCShrinkage function. Isoform names not conforming to ENSEMBL transcript names denote novel isoforms, and gene names not conforming to ENSEMBL gene names correspond to novel gene loci. Table 2 corresponds to results from gene set enrichment using gene names from used isoforms.

Supplemental Table 6 - List of differentially expressed genes and FLAIR isoforms. Same as Supplemental Table 5, but for differentially expressed genes and isoforms. Table 1 corresponds to differentially expressed isoforms as determined by DESeq2. Table 2 corresponds to results from gene set enrichment using gene names from differentially expressed isoforms. Table 3 corresponds to differentially expressed genes.

Supplemental File 1 - GTF of FLAIR isoforms. General transfer formatted (GTF) list of mRNA isoforms identified by FLAIR.

Supplemental File 2 - GTF of StringTie isoforms. Same as Supplemental File 1 but for isoforms assembled by StringTie.

Supplemental File 3 - FLAIR isoform count table. Tab delimited file containing raw expression from FLAIR-quantify. Columns correspond to isoform name and sample ids.