University of California
Santa Cruz

Alternative RNA Processing in Cancers

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

DOCTOR OF PHILOSOPHY

in

BIOMOLECULAR ENGINEERING AND BIOINFORMATICS

by

Alison D. Tang

June 2022

The Dissertation of Alison D. Tang is
approved:

_____
Professor Angela Brooks, Chair

_____
Professor Christopher Vollmers

_____
Professor Manuel Ares

_____
Peter Biehl
Vice Provost and Dean of Graduate Studies

Table of Contents

List of figures

Alternative RNA Processing in Cancers
Alison D. Tang
Abstract

RNA-Seq has brought forth significant discoveries concerning aberrations in RNA processing, implicating RNA variants in a variety of diseases. Here, I will discuss cancer-associated RNA variation accompanying either somatic mutations in a splicing factor or changes in RNA base editor abundance. While many splice variants have been examined at an event-level with short reads, identifying full-length isoform changes may better elucidate the functional consequences of these variants in cancer. Thus, we have employed long-read technology to obtain full-length transcript sequences, developing a computational workflow called FLAIR (Full-length Alternative Isoform analysis of RNA) to identify high-confidence transcripts, differential transcript usage, and haplotype-specific transcripts. We performed nanopore sequencing of chronic lymphocytic leukemia patient samples containing *SF3B1* mutation. With FLAIR, we are able to find patterns of aberrant splicing and a decrease in unproductive retained introns associated with *SF3B1* mutation. Additionally, we have applied FLAIR to direct RNA sequencing reads to facilitate the identification of longer poly(A) tail lengths associated with intron retentions. Finally, we have sequenced H1975 lung adenocarcinoma cells with knockdown of ADAR, an enzyme that mediates A-to-I editing. We further improved our workflow to identify key inosine-isoform associations with the goal of clarifying the prominence of ADAR in tumorigenesis. Ultimately, our work demonstrates the utility of nanopore sequencing for augmenting cancer and splicing research.

Acknowledgements

The text of this dissertation includes revised text of the following previously published material: Full-length transcript characterization of *SF3B1* mutation in chronic lymphocytic leukemia reveals downregulation of retained introns (Tang et al. 2020) and Nanopore native RNA sequencing of a human poly(A) transcriptome (Workman et al. 2019). My scientific achievements thus far and into the future are not my own but the product of so many. Thus, I owe credit to all the other brilliant scientists (*e.g.* the RNA Consortium) I have collaborated with to author these papers, all of whom have contributed to the work I describe here.

I thank my advisor Dr. Angela Brooks for embodying what modern and future scientists should be like. I am thankful for having her stellar example that I can only hope to emulate. For example, her supportiveness: she lifts up underrepresented minorites in science; is patient in my journey to being a better scientist; and has often encouraged me to seek and provided me the resources to have quintessential experiences in science (*e.g.* conferences, collaborations, mentees), building up my confidence and all with a pressure applied so kindly that even I could take it. I appreciate her for trusting me with a project as a rotation student and taking me in, especially because I most likely did not give off very promising vibes in the beginning. I also appreciate her style of leadership, which allows for this two-way transparency and for those under her to be seen and heard, all while still serving as our northern light with her intelligence and experience. With this non-exhaustive list, I conclude that Angela is a one-of-a-kind mentor to come across in science and I am fortunate enough to have been her student.

I would also like to acknowledge my family and friends, most of whom are not in science, who took the time to support me all the same. I would also like to thank the community of scientists I have gotten to know in my time as a graduate student. It's a large community – my lab mates; graduate student friends; folks in and PIs of the Vollmers, Ares, Carpenter, Arriberre, Akeson, and Klein labs; my thesis committee and those that sat on my advancement committee; and so many others that don't fall into those bins that have all impacted this journey so positively.

Alternative RNA Processing in Cancers

Introduction

RNA splicing is a carefully-regulated and a pervasive form of gene processing, with over 95% of human multi-exon genes being alternatively spliced (Pan et al. 2008; E. T. Wang et al. 2008). Unsurprisingly, splicing dysregulation is a recently appreciated hallmark of cancer (Garraway and Lander 2013; Oltean and Bates 2014). Studies have shown increased alternative splicing (AS) in tumors (Kahles et al. 2018), such as elevated levels of intron retention events (Dvinge and Bradley 2015), finding thousands of novel, tumor-specific AS. Furthermore, tumor-specific AS can result in the production of tumor-specific aberrant proteins (Kahles et al. 2018), demonstrating a mechanism via the translation of AS transcripts for the potential functional impact of splicing in disease. One form of AS called intron retention can result in the production of neoantigens, worsening multiple myeloma patient outcomes while also providing a therapeutic strategy for targeting multiple myeloma cells to prevent immune escape (Dong et al. 2021). The inhibition via small molecule of a specific splice variant of the *IRAK4* gene, shown to be present in a subset of acute myeloid leukemia and myelodysplastic syndrome patients, is tied to the abrogation of leukemic growth (M. A. Smith et al. 2019). AS and disease are further connected and resolved through the application of an antisense oligonucleotide to induce splice-switching in spinal muscular atrophy patients (Wurster and Ludolph 2018). With these examples and many more to be discovered, it is crucial that studies measure AS accurately as this may lead to additional insights into future prognosis and treatment of cancers and other genetic diseases.

The advent of Next Generation Sequencing has accelerated the discovery of alternative splicing breakthroughs in cancers and is correlated with a growth in research focused on splicing (Cloonan et al. 2008; Sultan et al. 2008). Short-read RNA-Seq, or shotgun sequencing, can survey complete transcriptomes with high depths. The general steps for short-read RNA-Seq are as follows: RNA transcripts are isolated, reverse transcribed to cDNA copies,

fragmented, and then sequenced using sequencing-by-synthesis techniques. The sequenced fragments range from 50-500 bp, limited mainly by the dephasing that can occur as lengths are increased (Nakamura et al. 2011). With RNA-Seq, transcriptome-wide characterization of AS in cancer became routine (L. Wang et al. 2016; Darman et al. 2015). Analyses of The Cancer Genome Atlas (TCGA) data, a Pan-Cancer repository profiling upwards of 20,000 cancer and matched normal samples with multiomic assays, has revealed mutations in splicing factor genes to be positively selected for in tumors of multiple cancer types (Kandoth et al. 2013; Seiler et al. 2018). RNA-Seq studies have also informed that higher rates of splicing factor mutations cause global splicing deregulation, suggesting that broad transcriptional alterations may benefit tumorigenesis (Seiler et al. 2018).

However, the resolution of alternative splicing events using short-read sequencing is restricted to characterizing the transcriptome at the level of distinct RNA processing events. As short read data are only capturing fragments of cancer transcripts, this necessitates a number of assumptions and inferences in order to reconstruct and quantify isoforms. For this reason, fragmented data lack the ability to identify the full transcriptional context of aberrant splice events, further convoluting the functional consequences of cancer-associated splicing events identified with short read RNA-Seq. The detection of coordinated splicing variation and RNA sequence variation is limited to the length of fragments which are sequenced (Steijger et al. 2013). Moreover, detection and quantification of transcripts containing retained introns using short reads is difficult and often overlooked (Jacob and Smith 2017; Q. Wang and Rio 2018). Other forms of RNA processing, such as RNA modifications, cannot be detected with short reads without modification-specific protocols (Lovejoy, Riordan, and Brown 2014; Carlile et al. 2014; Dominissini et al. 2013). RNA editing, while detectable computationally as in the case with A-to-I editing, can be challenging to detect in cases of hyper-editing (Porath, Carmi, and Levanon 2014). As with RNA splicing, RNA modifications and RNA edits are tied to disease; for example, the dysregulation of N6-methyladenine, the most common internal modification

on mRNA (N. Liu and Pan 2016), and has been linked to human diseases including obesity and cancer (Sibbritt, Patel, and Preiss 2013). Taken together, proper study of aberrations in co-transcriptional and post-transcriptional RNA processes individually and in combination with each other motivates the exploration of short read RNA-Seq alternatives.

Studying cancer transcriptomes at the level of individual variants limits our complete understanding of the functional consequences of these aberrations. Thus, the exploration of the unique application of long-read nanopore RNA-Seq to cancer transcriptomes is imperative for potentially overcoming Illumina length limitations. Long-read sequencing techniques offer increased information on exon connectivity by sequencing full-length transcript molecules (Bolisetty, Rajadinakaran, and Graveley 2015; Sharon et al. 2013; Bueno et al. 2016). We focused on nanopore sequencing, which works by measuring the change in electrical current caused by DNA or RNA threading through a nanopore and converting the signal into nucleotide sequences (Deamer, Akeson, and Branton 2016; Garalde et al. 2018). Nanopore sequencing yields long reads as long as 2 megabases (Payne et al. 2018) and has been used for applications ranging from the sequencing of the centromere of the Y chromosome (Jain, Olsen, et al. 2018), the human genome (Jain, Koren, et al. 2018) completely (Miga et al. 2020), and single-cell transcriptome sequencing (Byrne et al. 2017; Volden et al. 2018). Likely owing to the lower sequence accuracy of the reads (Deamer, Akeson, and Branton 2016), nanopore technology had yet to be thoroughly explored as a tool for detecting subtle splicing variation, nor had it been used to examine cancer-associated splicing factor mutations. In this dissertation, I discuss the incremental development of computational methods to analyze alternative RNA processing using nanopore reads, all in effort to improve our understanding of cancers built upon short read information.

Chapter 1: Resolving full-length isoforms in CLL patients with *SF3B1* mutation reveals changes in intron retention

1.1 Abstract

While splicing changes caused by somatic mutations in *SF3B1* are known, identifying full-length isoform changes may better elucidate the functional consequences of these mutations. We report nanopore sequencing of full-length cDNA from CLL samples with and without *SF3B1* mutation, as well as normal B cell samples, giving a total of 149 million pass reads. We present FLAIR (Full-Length Alternative Isoform analysis of RNA), a computational workflow to identify high-confidence transcripts, perform differential splicing event analysis, and differential isoform analysis. Using nanopore reads, we demonstrate differential 3' splice site changes associated with *SF3B1* mutation, agreeing with previous studies. We also observe a strong decrease in expression of intron retention events associated with *SF3B1* mutation. Full-length transcript analysis links multiple alternative splicing events together and allows for better estimates of the abundance of productive versus unproductive isoforms. Our work demonstrates the potential utility of nanopore sequencing for cancer and splicing research.

1.2 Introduction

In various cancers, mutations in the splicing factor *SF3B1* have been associated with characteristic alterations in splicing. In particular, recurrent somatic mutations in *SF3B1* have been linked to various diseases, including chronic lymphocytic leukemia (CLL) (L. Wang et al. 2011; Quesada et al. 2011; Rossi et al. 2011; Landau et al. 2015), uveal melanoma (Furney et al. 2013; Harbour et al. 2013; Martin et al. 2013), breast cancer (Maguire et al. 2015; Pereira et al. 2016; Fu et al. 2017), and myelodysplastic syndromes (Papaemmanuil et al. 2011; Malcovati et al. 2011). *SF3B1* is a core component of the U2 snRNP of the spliceosome and

associates with the U2 snRNA and branch point adenosine of the pre-mRNA (Gozani, Feld, and Reed 1996; Gozani, Potashkin, and Reed 1998; Will et al. 2001). Mutations in the HEAT-repeat domain of *SF3B1*, such as the K700E hotspot mutation, have been shown to be associated with poor clinical outcome in CLL (L. Wang et al. 2011; Quesada et al. 2011; Rossi et al. 2011; Landau et al. 2015).. B cell-restricted expression of *SF3B1* mutation together with *Atm* deletion leads to CLL-like disease at low penetrance in a mouse model, confirming a contributory driving role of mutated *SF3B1 (Yin et al. 2019)*. Additionally, mutations in *SF3B1* induce aberrant splicing patterns that have been well-characterized using short-read sequencing of the transcriptome. Most notably, mutant *SF3B1* has been shown to generate altered 3' splicing (L. Wang et al. 2016; DeBoever et al. 2015; Darman et al. 2015). Mutant *SF3B1*-associated changes in branch point recognition and usage (Carrocci et al. 2017; Kesarwani et al. 2017; Alsafadi et al. 2016) form the model in which mutant *SF3B1* affects acceptor splice sites. Targeting an aberrant branch point recognized by mutant *SF3B1* in the tumor suppressor *BRD9* has been shown to suppress tumor growth using antisense oligonucleotides (Inoue et al. 2019), further revealing the therapeutic implications for treating mutant *SF3B1*-induced mis-splicing.

To investigate $SF3B1^{K700E}$ AS at an isoform level with nanopore data, the representative transcripts and their abundance in each condition must be determined from the reads. Existing software for short-read RNA-Seq data that perform isoform assembly, splicing, or quantification analyses are not designed to work properly with the length of and frequent indels present in nanopore reads. The raw accuracy of nanopore 1D cDNA sequencing is approximately 85-87% (Jain et al. 2017; Volden et al. 2018; Workman et al. 2018), although accuracy can change depending on iterations of the technology and library preparation methods (Volden et al. 2018). To assemble isoforms and perform splicing analysis from nanopore reads, we have created a workflow called Full-Length Alternative Isoform analysis of RNA, or FLAIR. FLAIR requires a reference genome to define isoforms from long reads. While

FLAIR does not require short reads, having matched short-read data can be used to identify unannotated splice sites and improve the confidence of transcript splice junction boundaries. Recognizing the benefit of highly-accurate short reads for detecting the splice junctions of a mutated splicing factor, we used a hybrid-seq approach in this study. We combined the accuracy of Illumina short reads for splice junction accuracy with the exon connectivity information of long reads to overcome the higher error rates of long reads.

Of a large cohort of CLL patient tumor samples characterized using short-read RNA-Seq (L. Wang et al. 2016), we present the resequencing of a subset of these transcriptomes, globally, with nanopore technology: three with wild type *SF3B1*, three with the K700E mutation, and three additional normal B cell samples, which are the normal lineage cellular complement to CLL, to use as a normal tissue control (Wan and Wu 2013). Following the identification of high-confidence isoforms from nanopore data, FLAIR provides a framework for performing alternative splicing and differential isoform usage analyses. Upon splicing analysis of the nanopore CLL data, we observe a bias toward increased alternative 3' splice sites (3'SS) over alternative 5' splice sites (5'SS) in CLL $SF3B1^{K700E}$ samples, consistent with the known effects of *SF3B1* mutation. We also highlight a previously underappreciated finding of differential intron retention in CLL $SF3B1^{K700E}$ versus $SF3B1^{WT}$ with increased splicing relative to wild type *SF3B1* samples. Using long reads, we can identify retained introns more confidently than with short reads and are able to observe AS events across full-length isoforms. FLAIR analysis of nanopore data reveals new biological insights into *SF3B1* mutations and demonstrates the potential for discovering new cancer biology with long-read sequencing.

1.3 Sequencing summary

We resequenced six primary CLL samples and three B cells, generating 257 million total reads with large variability in read depth and percentage of pass reads for each flow cell (Table 1). On average, 30.5% of the PromethION reads were considered full-length (Methods).

1.4 Developing and benchmarking FLAIR

We developed FLAIR to generate a set of high-confidence isoforms that were expressed in our samples. FLAIR summarizes nanopore reads into isoforms in three main steps: alignment, correction, and collapse (Fig. 1.1). In the alignment step, we aligned raw read sequences from all samples to the genome to identify the general transcript structure. We compared the long-read spliced-aligners minimap2 (Li 2016) and GMAP (Wu and Watanabe 2005), the latter of which has been used in several other long-read studies (Byrne et al. 2017; Weirather et al. 2017; Križanovic et al. 2018). The aligners were evaluated on a subset of the CLL data according to splice-site accuracy, comparing splice sites mapped by each aligner with annotated splice sites. Minimap2 demonstrated marked improvement in splice-site mapping over GMAP (not shown).

The next main step of FLAIR is to *correct* splice sites because the relatively high-sequencing error rates, frequent base deletions, and difficulties of spliced alignment can result in spurious alignments near splice sites. To address indels, small gaps in the read alignments were artificially filled (Methods). FLAIR only considers a splice site as "correct" as long as it is supported by orthogonal data, such as splice sites curated in annotations or observed in matched short read sequencing. To correct splice sites, an incorrect splice site in a read alignment is replaced with a correct one as long as the correct splice site is within a window size of 10 bp. We initially identified many minimap2-aligned splice sites that aligned outside of the window size from the nearest supported splice site, were unannotated, and had no short-read support; we determined that these were of lower confidence based upon further manual inspection. Many of these novel sites appeared to be driven by alignment errors. Thus, we did not consider nanopore reads with novel splice sites that had no additional support.

The *collapse* step of FLAIR produces a set of expressed isoforms with high confidence for further downstream analysis. Starting with only the fully splice-corrected reads, FLAIR

constructs a first-pass nanopore isoform transcriptome, collecting the reads with the same unique splice junctions chain into distinct isoform groups (Fig. 1.1b). FLAIR determines one or more representative isoform(s) within each group by calling confident transcription start and end sites based on the density of read start and end positions (Methods). Next, FLAIR reassigns all of the reads to a first-pass isoform, including reads with minimap2-aligned splice sites that were not able to be fully splice corrected. This is done by aligning the reads to spliced sequences from the first-pass isoform set (*i.e.,* aligning reads to an isoform sequence without a spliced alignment) and assigning the read to an isoform with the best alignment with MAPQ ≥ 1. This realignment of reads to the set of nanopore-specific isoforms accounted for misalignments that manifested from spliced alignment by constraining reads to align only to splice junction chains with additional support. The realignment was also crucial for better distinguishing splice-site differences (L. Wang et al. 2016). Finally, FLAIR filters out first-pass isoforms that have fewer than three supporting reads and the remaining isoforms with sufficient coverage comprises the final high-confidence nanopore isoform set.

1.5 CLL isoforms detected from FLAIR

Using FLAIR, we identified a total of 326,699 high-confidence spliced isoforms. Of these isoforms, 32,479 matched annotated isoforms and the majority (90.0%) were unannotated. Most of the unannotated isoforms were a novel combination of already annotated splice junctions (142,971), while others deviated from the annotation because they contained a retained intron (21,700) or a novel exon (3594). The remainder of the unannotated isoforms contain at least one novel splice site not present in annotations but supported through short reads. We performed a saturation analysis of the reads, by condition, to assess the number of FLAIR isoforms that could be detected at differing read depths. For all conditions, we found that sequencing at greater depth would facilitate the discovery of even more isoforms. Predictably, at comparable read depths, we were able to detect the greatest number of isoforms in CLL *SF3B1*^{K700E}. The read lengths of CLL *SF3B1*^{WT} were noticeably shorter, resulting in

fewer isoforms able to be detected compared with B cell or CLL *SF3B1*^{K700E}. The saturation analysis illustrates the diversity of isoforms resulting from mutant *SF3B1* as well as the importance of read length for isoform discovery.

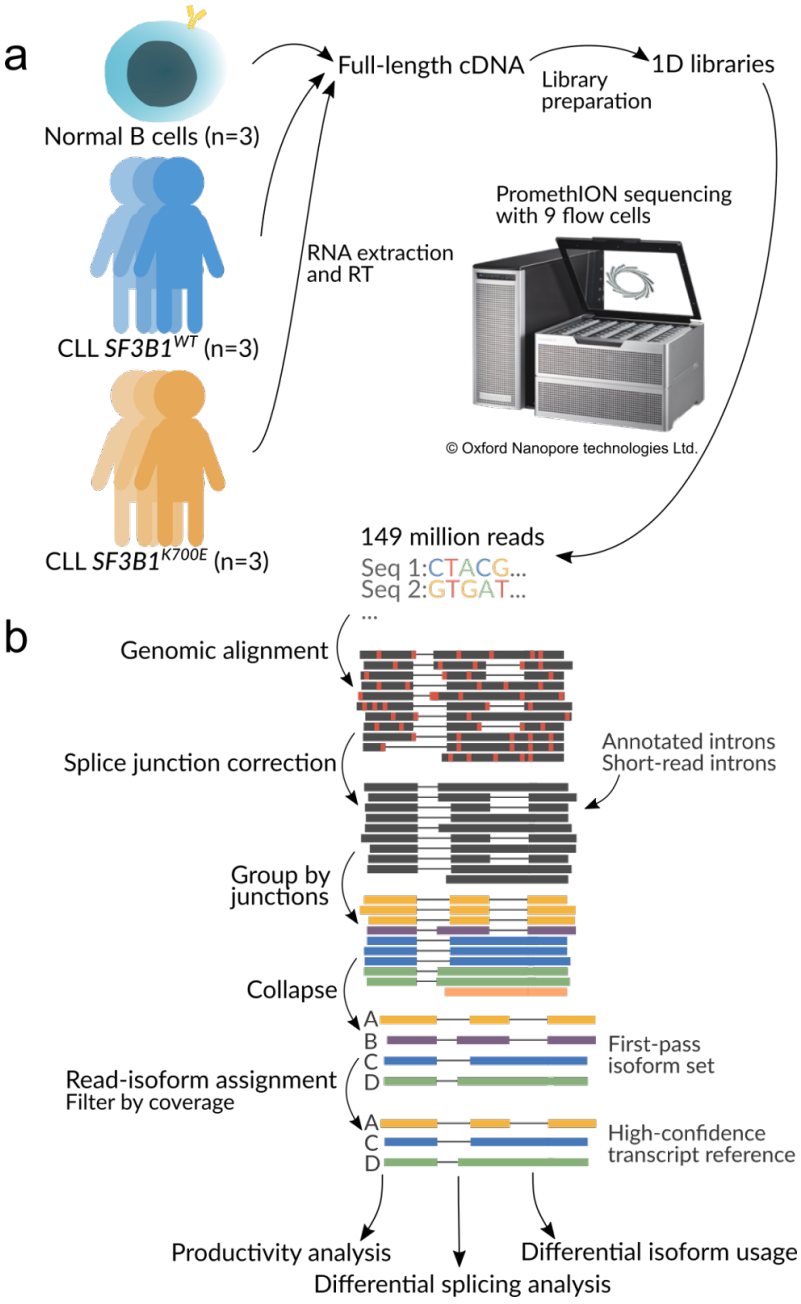

Figure 1.1 Long-read nanopore sequencing and FLAIR analysis to identify full-length transcripts associated with SF3B1 mutation in chronic lymphocytic leukemia. a *RNA from primary samples across three conditions (chronic lymphocytic leukemia with and without SF3B1 mutation and normal B cells) were obtained. The RNA was prepared into 1D cDNA libraries and each sample was sequenced on a PromethION flow cell. The basecalled data were processed using the FLAIR pipeline.* ***b*** *The FLAIR pipeline constructs an isoform set from nanopore reads. First, reads are aligned to the genome with a spliced aligner. The sequence errors are marked in red. Next, they are splice-corrected using splice sites from either annotated introns, introns from short-read data, or both. The corrected reads are grouped by their splice junction chains and are summarized into representative isoforms (first-pass set). All reads are then reassigned to a first-pass isoform. The isoforms that surpass a supporting read threshold of 3 comprise the final high-confidence isoform set.*

1.6 Comparison of short- and long- read alternative 3' splicing analysis

Previous studies have demonstrated that *SF3B1*^{K700E} promotes alternative 3′ splice-site (3'SS) usage (L. Wang et al. 2016), a pattern we sought to validate in our nanopore data. From the cohort of 37 CLL samples (L. Wang et al. 2016) (24 *SF3B1*^{WT} and 13 with *SF3B1* mutation) sequenced with Illumina short reads, 65 significantly altered 3′ splicing events associated with *SF3B1*^{K700E} were identified using juncBASE (Brooks et al. 2011) For those significant events, we measured the change in percent spliced-in (dPSI) values using the corrected nanopore reads by subtracting the PSI of CLL *SF3B1*^{WT} from the PSI of CLL *SF3B1*^{K700E} and compared them with the short-read dPSIs (Fig. 1.2a). The long-read dPSIs were correlated with the CLL short-read dPSIs, and dPSIs were more similar across the two sequencing technologies with increasing long-read depth (Pearson correlation 0.952). Some splice junctions had insufficient coverage in our nanopore data for adequate power to detect the same splice-site usage observed with short reads. Additionally, we wanted to identify the altered 3′ splicing events that would be significantly altered by mutant *SF3B1* in the nanopore data alone.

We called alternative 3′ and 5′ splice sites observed in the FLAIR isoforms and quantified the coverage of the alternative events in each of the CLL samples. We identified 35 alternative 3′SSs and 10 alternative 5′SSs that were significantly differentially spliced (corrected p value < 0.1 and dPSI absolute value > 10) between $SF3B1^{K700E}$ and $SF3B1^{WT}$. More $SF3B1^{K700E}$-associated 3′ alterations were upstream of $SF3B1^{WT}$-associated 3′SSs (20 out of 35) and only 2 of the 35 alternative 3′SSs had been previously identified with short-read sequencing.

The distribution of distances between $SF3B1^{K700E}$-altered 3′SSs to canonical sites peaks around −20 bp and is significantly different from a control distribution (two-sided Mann–Whitney U $p = 6.77 \times 10{-}2$) (Fig. 1.2b), similar to what has been reported in CLL short-read sequencing (L. Wang et al. 2016). We were unable to find any unifying sequence motif associated with these altered 3′SS identified in the nanopore data. However, using the 65 alternative 3′SSs significantly associated with *SF3B1* mutation identified in the CLL short-read data, we found a tract of As 13–16 bp upstream of the canonical 3′SS. This motif is concordant with other mutant *SF3B1* studies using short reads (DeBoever et al. 2015; Darman et al. 2015).

One of the alternative 3′SS identified from both long and short reads was in the *ERGIC3* gene (Fig. 1.2c). There were two dominant isoforms: a novel isoform containing the proximal splice site that was more abundant in $SF3B1^{K700E}$ and another annotated isoform containing the distal splice that was expressed in both the mutant and wild type samples. Both the proximal and distal 3′SS were associated with multiple isoforms with distinct AS patterns up- and downstream of the alternative 3′SS. Long reads enabled us to not only identify mutant *SF3B1*-altered splice sites, but also associate an event-level aberration with full-length isoforms containing other alternative processing events.

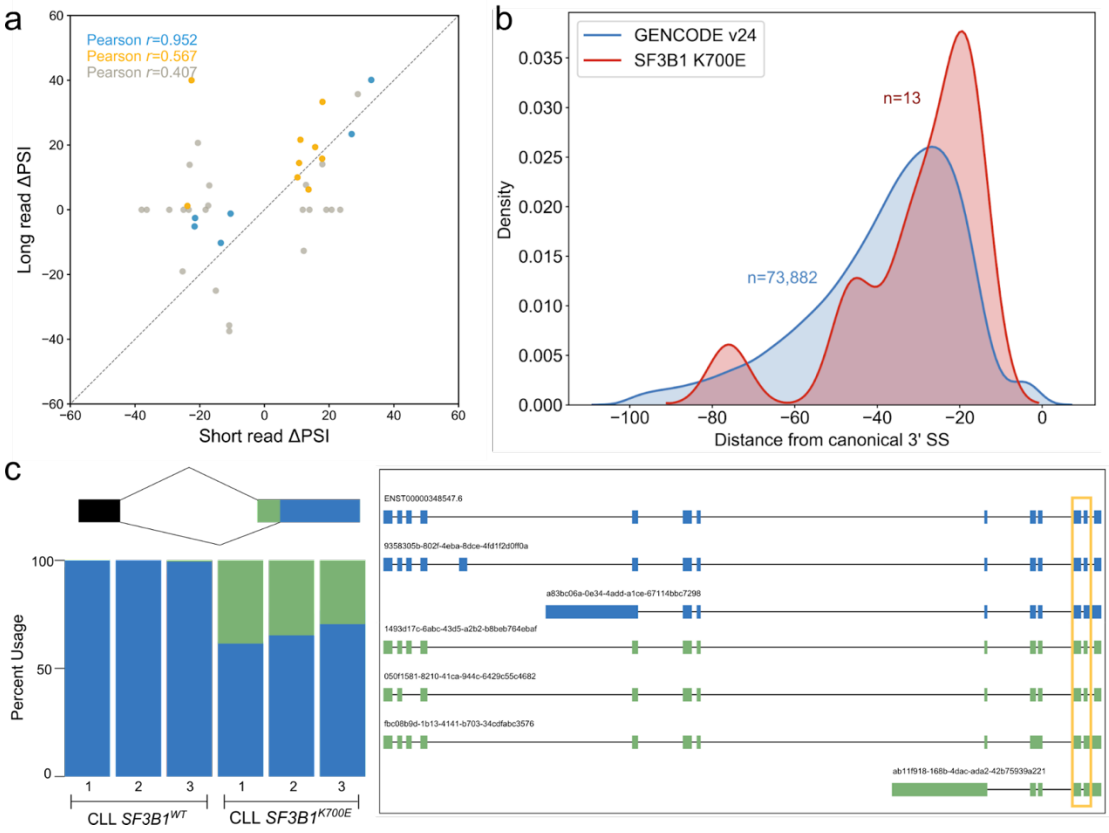

Figure 1.2 Evaluation of nanopore read alignment, correction, transcript assembly, and transcript quantification. ***a*** *Comparison of the delta PSI for significant alternative splicing events identified by juncBASE using short-read data and the median delta PSI of the same events using long-read data. The colors correspond to the median coverage of the splice junctions, where blue is greater than 25 reads, yellow is greater than 10, and gray is greater than 0.* ***b*** *Distribution of proximal splice sites found in nanopore reads for 13 significant sites (corrected p value < 0.10, delta PSI > 10). The GENCODE distribution is the distribution of distances from canonical 3' splice sites to the nearest non-GAG trimer.* ***c*** *ERGIC3 splice-site usage (left) and full isoforms (right) for the proximal chr20:35,556,954 (green) and distal chr20:35,556,972 (blue) sites from 5' to 3'. The dominant isoforms using either the proximal or distal site are 1493d17c-6abc and ENST00000348547.6. The alternative acceptor event is boxed in yellow in the isoform schematic.*

1.7 Intron retention events have decreased expression in CLL SF3B1^{K700E}

Intron retentions (IR) have been observed to differentiate tumors from matched normal tissue, as they are highly prevalent across a variety of cancers (Dvinge and Bradley 2015; Jung et al. 2015). However, based on common approaches used by short-read AS analysis tools, it is difficult to characterize IR event usage confidently using short reads (Jacob and Smith 2017). Thus, unless stringent approaches are used, intron retention events are easily misclassified particularly in regions with complex AS (Q. Wang and Rio 2018). With long reads, a single read is capable of connecting multiple AS events in addition to spanning IR, enabling easier identification and quantification of IR. To investigate changes in IR associated with *SF3B1*^{K700E}, we categorized each FLAIR isoform as IR-containing or not (spliced). Comparing the expression fold-change between CLL samples revealed that IR isoforms were globally downregulated in the *SF3B1*^{K700E} sample compared with CLL *SF3B1*^{WT} (Fig. 1.3a). When performing the same comparison between B cell and *SF3B1*^{K700E}, we observed no significant difference in the expression of IR-containing isoforms (two-sided Mann–Whitney U $p = 0.121$). Reanalysis of the CLL short-read data confirmed the observed increase in the inclusion of retained introns in CLL *SF3B1*^{WT} samples (Fig. 1.3b).

To further investigate the effect of *SF3B1*^{K700E} on increased intron splicing, we reanalyzed Nalm-6 Pre-B isogenic cell lines (Darman et al. 2015) with either *SF3B1*^{WT} or *SF3B1*^{K700E} sequenced using short reads. We used juncBASE (Brooks et al. 2011) to identify and quantify AS between the two conditions. For the 16 significant (corrected $p < 0.1$) IR events, Nalm-6 *SF3B1*^{K700E} PSI values appeared lower than *SF3B1*^{WT} (Fig. 1.4a), supporting a decrease in retained introns in *SF3B1*^{K700E}-containing samples; however, the difference was not significant. We observed that for IRs that were more spliced with mutant *SF3B1*, they were spliced with greater magnitude than the IRs more spliced in the wild type (Fig. 1.4b). In addition, we reanalyzed TCGA breast cancer samples without common splicing factor mutations against samples with *SF3B1*^{K700E} using juncBASE and found the same trend of increased IR splicing in *SF3B1*^{K700E} (Fig. 1.4c).

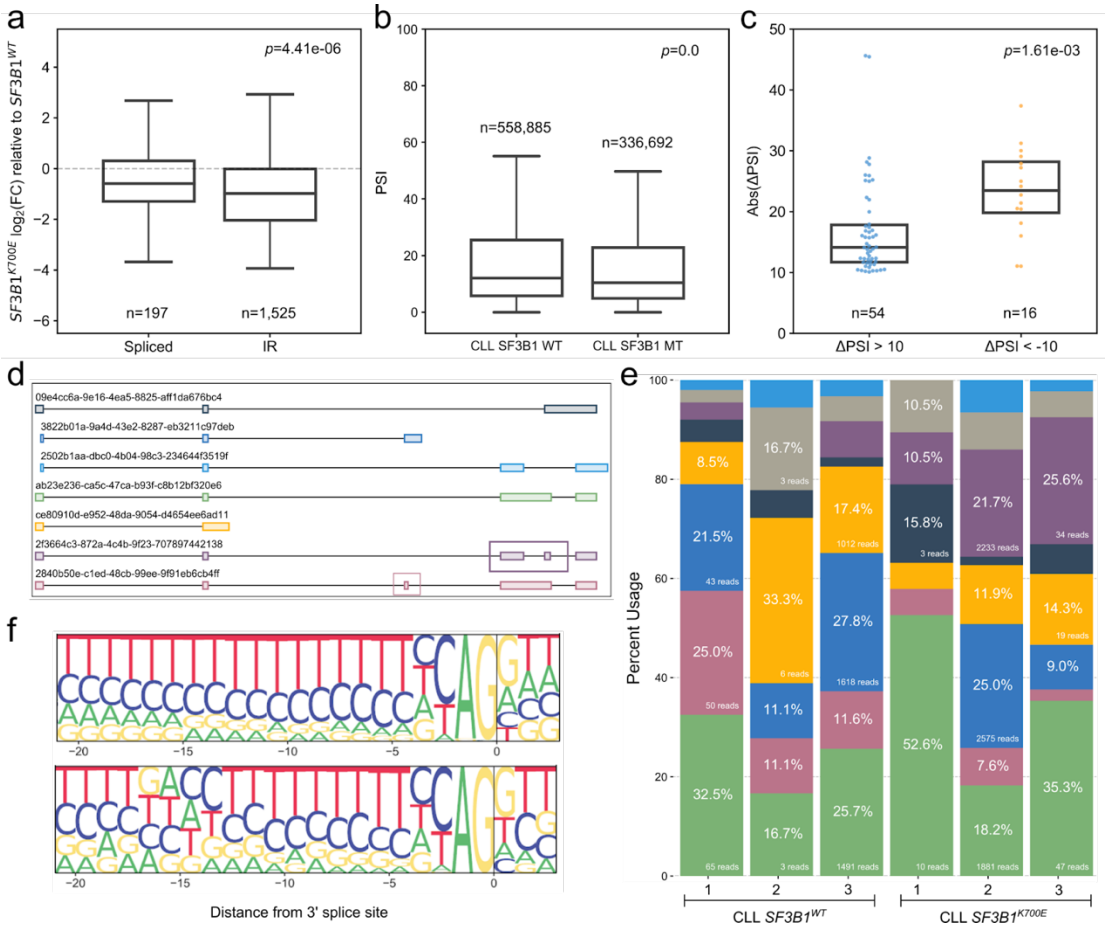

Fig 1.3 Intron retention events are expressed significantly less in CLL SF3B1^{K700E}. ***a*** *Expression fold-change between SF3B1K700E and SF3B1WT of FLAIR isoforms with (IR) or without (spliced) retained introns. Boxplot median difference = 0.395.* ***b*** *PSI values of intron retention events identified in short read sequencing of CLL SF3B1 WT or CLL SF3B1 MT samples. Boxplot median difference = 1.69.* ***c*** *The change in PSI in significant intron retention events (corrected p < 0.1) identified in nanopore data that are more included in CLL SF3B1K700E (blue) or more included in CLL SF3B1WT (orange). Boxplot median difference = 9.32. P values for a–c are using two-sided Mann–Whitney U tests. Box-plots show median line, box limits are upper and lower quartile, and whiskers are 1.5× interquartile.* ***d*** *ADTRP gene isoforms, plotted 5' to 3'. The 632 bp intron that is differentially included is boxed in purple. A differentially skipped exon is boxed in pink.* ***e*** *Percent usage of each isoform in each CLL sample, with colors corresponding to isoforms in* ***d****. Gray bars represent all other isoforms not plotted in* ***d****.* ***f*** *Top: 3' splice-*

site motif of constitutively spliced introns. Bottom: 3' splice-site motif of significant intron retention events identified from short-read sequencing (n = 67).

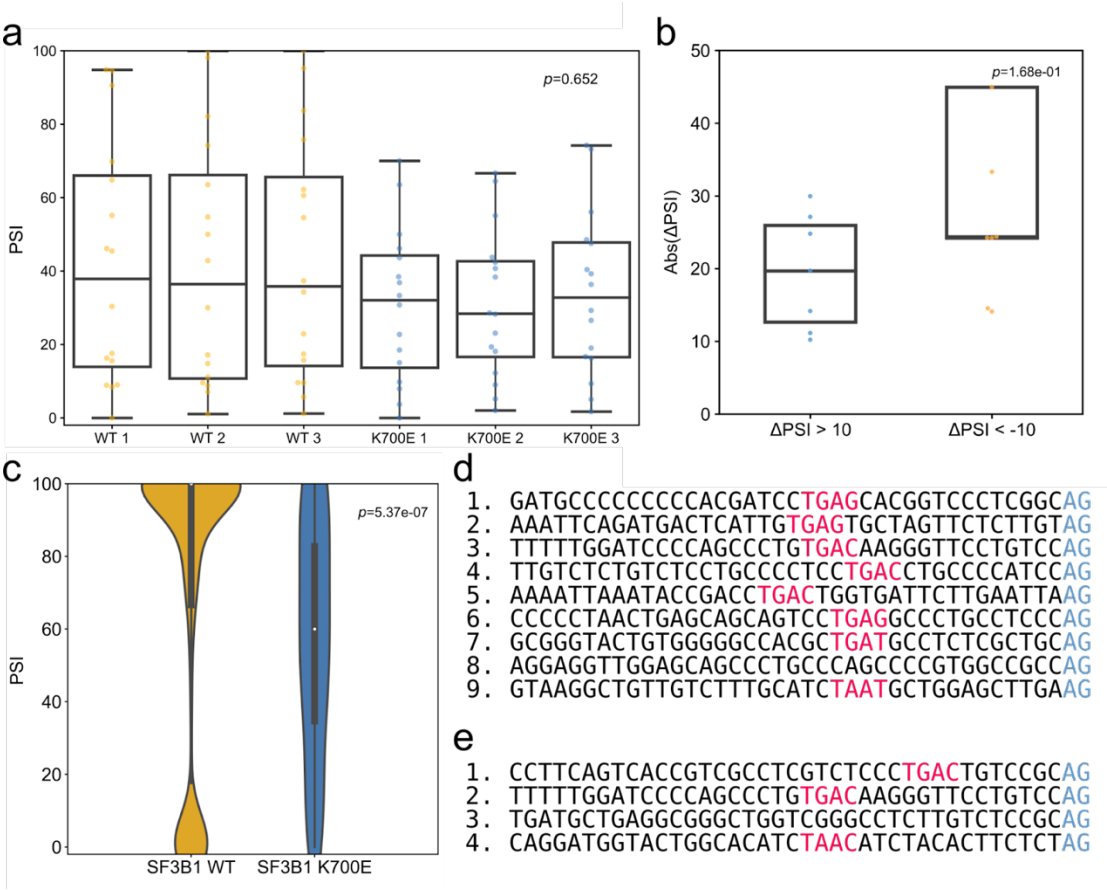

Fig 1.4 IR analysis of short read RNA-Seq data of Nalm-6 and TCGA BRCA samples with SF3B1^{K700E}. ***a*** *PSIs for 16 significant (corrected p < 0.1) IR events in 6 Nalm-6 samples, 3 with wildtype SF3B1 and 3 with SF3B1^{K700E}. The P-value is calculated from a Kruskal-Wallis H test.* ***b*** *The change in PSI in significant intron retention events (corrected p < 0.1) identified in the Nalm-6 data that are more included in CLL SF3B1^{K700E} (blue) or more included in CLL SF3B1^{WT} (orange). Box-plots show median line, box limits are upper and lower quartile, and whiskers are 1.5x interquartile.* ***c*** *5 significant IR events were associated with SF3B1^{K700E} mutation in TCGA BRCA samples. The violin plots are made from individual PSIs for these IR events from: (SF3B1 WT) 801 samples without common splicing factor mutations and (SF3B1^{K700E}) 13 samples with SF3B1^{K700E}. Plots show median as white dot, box limits are upper and lower quartile, and filled area represents the entire range of the kernel density estimation. P-*

value is from a two-sided Mann-Whitney U test. ***d*** *3' splice site sequences for the 9 significant IRs from the Nalm-6 analysis that were more included in the WT. Red: motifs that are similar to the branch point motif in Corvelo et al. 2010; yellow: 3' splice site AG dinucleotide.* ***e*** *3' splice site sequences for the 4 significant IR events that were more included in the WT identified in the TCGA BRCA samples.*

Seeing that the trend of higher IR expression in *SF3B1*^{WT} was observed transcriptome-wide in the nanopore data, we narrowed our focus to only the introns that were significantly differentially retained between *SF3B1*^{K700E} and *SF3B1*^{WT}. Using DRIMSeq (Nowicka and Robinson 2016) for testing the IR events we identified from FLAIR isoforms, we found 70 introns were significantly different (corrected $p < 0.1$ and abs(dPSI) >10) with no overlap between these nanopore-identified events and the Illumina-identified events. Although there were fewer significant introns found to be downregulated in the mutant (dPSI <−10), the magnitude of downregulation was stronger for those introns (Fig. 1.3c). An example of a gene with increased expression of an isoform with increased IR splicing in CLL *SF3B1*^{K700E} is *ADTRP* (androgen dependent TFPI regulating protein), which is involved with blood coagulation (Lupu et al. 2011). We identified coordinated splicing events in *ADTRP*, such as isoforms with a differentially skipped exon coordinated with the differentially retained intron (Fig. 1.3d,e).

1.8 Strong branchpoint sequence identified associated with downregulated IR

We then looked at splice-site motifs for the differentially retained introns (Fig. 1.3f). In the IR events identified from the CLL short reads where the IR was more spliced in the mutant *SF3B1* condition, we found a strong TGAC branch point motif (Corvelo et al. 2010) 15 bp upstream of the 3′SS (Fig. 1.3f). While this motif had not been reported before in this context, it was consistent with the position of strong branch point sequences upstream of alternative 3′SSs that were associated with *SF3B1* mutation (Alsafadi et al. 2016). Sequence analysis of introns with increased inclusion in *SF3B1*^{WT} identified from both the Nalm-6 and TCGA BRCA short-read data also revealed a TGAC/TGAG motif ~15 bp upstream of the 3′SS (Fig. 1.4d,e), although not at exactly the same position. This further supports an underlying mechanism of

SF3B1^{K700E} in which the mutant prefers splicing at a 3′SS ~15 bp downstream of a strong branch point (Alsafadi et al. 2016). We did not observe the same motifs for the IR events identified from nanopore sequencing. To further investigate differences between IR events identified from nanopore sequencing compared with short-read sequencing, we looked at the intron length distributions. The median read lengths for nanopore reads were 712–948 bp, suggesting a bias against detecting longer IR. Indeed, the majority of differential IR identified in the nanopore data were under 1,000 bp, much shorter compared with those identified from short reads (Fig. 1.5). Thus, while we were able to identify a strong branch point sequence associated with IRs in several short-read datasets, we were unable to do so in the long reads in part because of a length bias in nanopore reads.

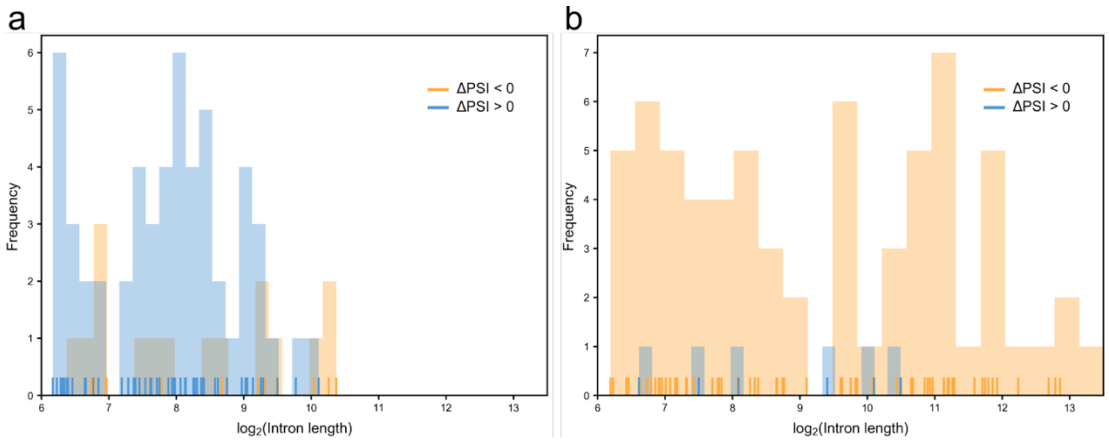

Fig 1.5 Nanopore and short-read differentially retained intron lengths. a *Histogram of the lengths of significant intron retention (IR) events between SF3B1*^{WT} *and SF3B1*^{K700E} *identified in the long-read data. The ticks along the x-axis are the individual intron lengths. Orange, IR events more included in the wildtype. Blue, IR events more included in the mutant.* ***b*** *Histogram of the lengths of significant IR events between mutant and wildtype SF3B1 identified from short-read data. The coloring is the same as in* ***a****.*

1.9 SF3B1^{K700E} *downregulates unproductive intron retention*

Short-read studies have noted an association between mutant *SF3B1* in CLL and an increase in transcripts with computationally predicted premature termination codons (PTCs)

(Darman et al. 2015). With full-length cDNA sequencing, we are given a more accurate representation of the complete transcript and thus are better able to detect transcripts with PTCs and estimate the proportion of unproductive transcripts. Unproductive isoforms are defined as those that have a PTC 55 nucleotides or more upstream of the 3' most splice junction (Rivas et al. 2015; Lewis, Green, and Brenner 2003) (Figure 1.6a). Productive transcripts are presumed to be protein-coding, whereas unproductive transcripts are either detained in the nucleus or subject to nonsense-mediated decay (NMD) if exported to the cytoplasm (Lewis, Green, and Brenner 2003; Sun et al. 2010; Filichkin and Mockler 2012; X.-D. Fu 2017). For example, *SRSF1* has several unproductive transcripts that are known to be either nuclear-retained or NMD-triggering (Sun et al. 2010), two of which (ENST00000581979.5 and unannotated Isoform V (Sun et al. 2010)) we were able to identify and accurately predict as unproductive in our nanopore data. We also identified 5 additional unannotated *SRSF1* isoforms with more than 100 supporting reads, 2 of which are productive and 3 are unproductive.

1.10 GO analysis of IR genes with decreased expression

Although together productive and unproductive IR isoforms were expressed less in *SF3B1*^{K700E} (Fig. 1.3a), the reduction was more pronounced in the unproductive IR isoforms (Fig 1.6b, productive-spliced and unproductive-IR two-sided Mann–Whitney U $p = 1.25 \times 10{-}6$). To further understand the decrease in unproductive IR observed in *SF3B1*^{K700E}, we performed a gene ontology (GO) analysis of the parent genes for the 94 isoforms in that category. No category reached statistical significance (corrected $p < 0.05$); however, the most enriched terms included antigen processing and presentation, cell cycle, regulation of MAP kinase activity, and positive regulation by protein kinase activity. The prevalence of cellular signaling GO terms parallels a finding in glioblastoma, where genes with a decrease in detained introns

regulated by *PRMT5* are also associated with perturbed kinase signaling (Braun et al. 2017).

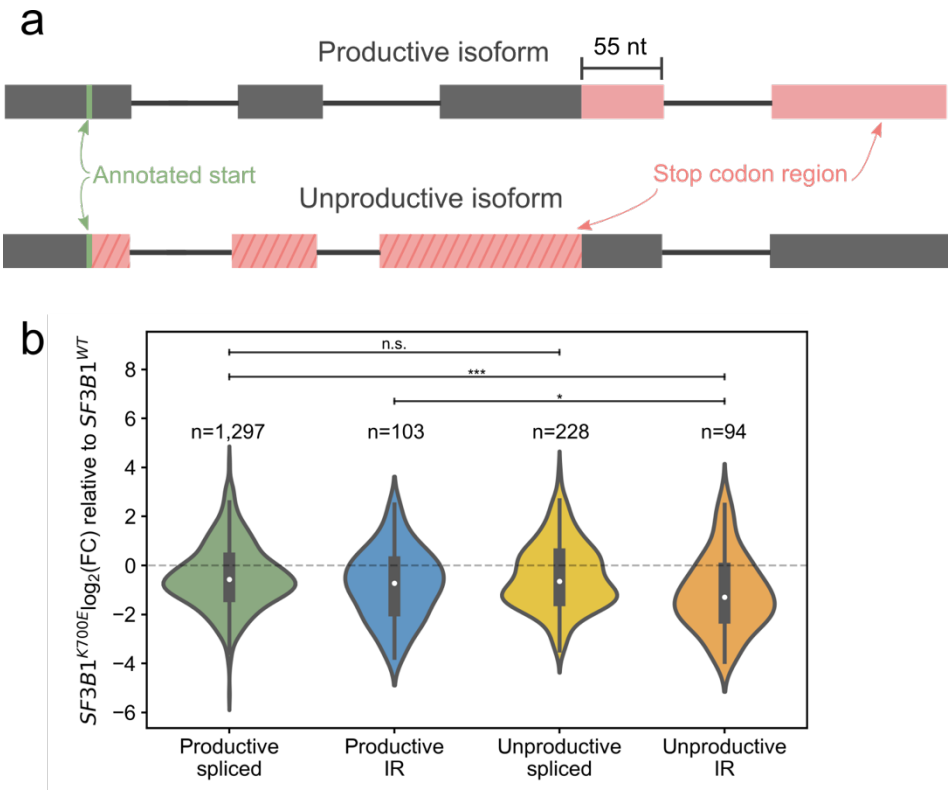

Figure 1.6 Mutant SF3B1 downregulates unproductive, intron-retaining transcripts. a *Schematic of productive and unproductive isoforms. The region on an isoform where stop codons can occur is colored red. Unproductive isoforms have premature stop codons present 55 nt or more from the last splice junction.* ***b*** *Expression fold-change (FC) between SF3B1^{K700E} and SF3B1^{WT} of FLAIR isoforms categorized as containing (IR) or not containing (spliced) retained introns and by productivity. The difference between the log$_2$(FC) of the productive spliced and the unproductive-IR categories is 0.720. Violin plot show median as white dot, box limits are upper and lower quartile, and filled area represents the entire range of the kernel density estimation. *p value < 0.05, ***p value < 0.0005, two-sided Mann–Whitney U test.*

1.11 Discussion

We report decreased intron retention in *SF3B1^{K700E}* in several instances: in our long-read RNA-Seq of 6 CLL samples; in Illumina RNA-Seq of a cohort of 37 CLL samples; and in 6 NALM cells with *SF3B1^{K700E}*. We postulate that these downregulated IRs are detained introns,

which could serve as a reserve of transcripts in the nucleus for cells to harness at a moment's notice. A GO analysis of these introns reveals many kinase signaling genes with connections to cancer. RNA-Seq, however, merely captures a snapshot of the transcripts present in the cell at a given time. It is unable to reflect the rates in which RNAs are transcribed, spliced, and degraded. What may appear as increased splicing efficiency of *SF3B1*^{K700E} could be the result of a combination of these processes being altered. Future studies to elucidate the splicing rate of K700E or interrogate the nuclear transcripts associated with *SF3B1*^{K700E} would be necessary to address these questions.

In this study, we identified splicing changes in the context of full-length isoforms in primary CLL samples with and without a mutation in splicing factor *SF3B1*. We were able to achieve high sequencing depths for long-read sequencing standards using the nanopore PromethION. Across the nine samples with great flow cell to flow cell variability in sequencing depth, we were able to generate 149 million pass reads. The errors in nanopore reads pose a challenge for many existing tools, e.g. alignment artifacts posing as novel splice sites. We developed FLAIR, a tool for the identification of high-confidence full-length isoforms and quantification of alternative splicing in noisy long read data. With FLAIR splice correction using matched CLL short reads, we rescued reads with incorrect splice sites for further analysis. FLAIR then defined a high-confidence isoform set for the nanopore CLL data as follows: (1) the fully corrected reads were collapsed to define a first-pass isoform set with vetted splice junctions, (2) all of the reads were reassigned to an isoform to assist with quantification of the aforementioned isoform set, and (3) isoforms with insufficient support were removed from the isoform set. FLAIR demonstrates improvements over the sparse space of nanopore analysis tools and enabled the discovery of many novel, *SF3B1* mutant-associated high-confidence isoforms.

Using FLAIR-defined transcripts, we identified aberrant splice site and retained intron usage associated with *SF3B1*^{K700E}. The alternative 3'SS usage patterns were consistent with

alterations identified in short-read data. In addition, long-read sequencing highlighted an expression decrease of isoforms containing retained introns in *SF3B1*^{K700E} relative to *SF3B1*^{WT}. This decrease was corroborated by reanalyzing CLL, Nalm-6 cell lines, and TCGA BRCA short-read datasets with mutant *SF3B1.* CLL has been shown to contain elevated levels of splicing alterations, regardless of SF3B1 mutation status (Ten Hacken et al. 2018; Yin et al. 2019). The subset of introns that exhibit increased splicing in the mutant point to a different intron retention landscape in CLL *SF3B1*^{K700E}. Introns more significantly spliced out in *SF3B1* mutated samples contained a strong branch point TGAC sequence ~15 bp upstream of the 3'SS, consistent with previously reported branch site motifs of altered 3'SSs associated with the mutation (Alsafadi et al. 2016).

Full-length reads also allow for improved identification of intron retentions and classification of transcript productivity, improving our understanding of *SF3B1* biology in CLL. Most notably, we observed a decrease in expression of intron-retaining isoforms categorized as unproductive in mutant *SF3B1*. Previous publications with short-read sequencing have shown that *SF3B1* mutation causes lower expression of genes with unproductive isoforms (Darman et al. 2015). As short-read sequencing has greater depth, it is easier to detect unproductive transcripts, many of which can be lowly expressed due to NMD. We speculate that the more highly expressed unproductive transcripts we detected with nanopore sequencing are likely retained in the nucleus. Performing a gene ontology analyses revealed kinase signaling associated with the unproductive IR events with decreased expression in *SF3B1*^{K700E}. We postulate that these unproductive retained introns are cases of detained introns, as kinase signaling has been associated specifically with detained introns (Braun et al. 2017). The perceived downregulation of these unproductive detained introns may result in increased production of kinase signaling genes to support tumor proliferation. Further experimentation would be necessary to verify that the unproductive IR events are retained in

the nucleus (detained introns) and if there is a functional relationship between kinase signaling genes.

A subsampling analysis revealed that we have not saturated the number of discoverable isoforms. Despite efforts to obtain nanopore sequencing data from a more high-throughput sequencing platform (PromethION) and account for the low accuracy of 1D nanopore sequencing, we note that the read depth, cohort size, splicing complexity, and high error rates in nanopore data are still limiting factors of this study. While we were able to detect alternative 3'SSs recapitulating *SF3B1* biology, nanopore sequencing was not able to detect as many altered events as short-read sequencing potentially due to the smaller cohort and the difficulty of detecting subtle splicing alterations. The small overlap between nanopore-identified and short read-identified alternative 3'SSs could also be due to the stringent filtering applied in an effort to determine alternative splicing more accurately. In addition, we did not find a strong branch point motif near the 3'SS of nanopore-identified IR events. This may have been due to a smaller cohort size and the bias toward shorter retained introns (<1,000 bp) sequenced in long reads (Fig 1.5). Everything considered, studying splicing factor mutations in primary patient samples using nanopore sequencing with fewer reads than the current study or without short-read sequencing would be suboptimal. Short-read sequencing was necessary for increasing confidence in splice sites, although future work with higher accuracy reads could potentially obviate the need for short reads. Future studies of primary samples should also include larger cohort sizes, with three replicates being the minimum (Schurch et al. 2016). Even though short-read technology is able to sequence more deeply than long reads, the ability of short reads to saturate splice junction detection depth-dependent (Nellore et al. 2016); thus, splicing studies should aim to sequence as deeply as possible. Fortunately, the throughput and accuracy of nanopore and PacBio technology has the potential to increase with subsequent iterations of the technologies (Jain et al. 2016). For nanopore in particular, methods to achieve higher sequence accuracy (Volden et al. 2018) or circumvent PCR bias and reverse

transcription length restrictions (Garalde et al. 2018) have been developed. In line with the rigorous pace of improvements in the field of long reads, PacBio has recently improved their throughput 8X with the newest PacBio Sequel II system, which has been shown to generate ~19 and 83 Gb of consensus reads (Vollger et al. 2019; Kingan et al. 2019).

This study of six primary CLL samples with nanopore sequencing demonstrates the ability of the nanopore to identify and quantify cancer-specific transcript variants. Long reads enabled us to better identify IR events, better estimate isoform productivity, and observe AS complexity in full-length isoforms. Ultimately, nanopore sequencing facilitated the building of a more complete picture of the transcriptome in primary cancer samples. With the impending rapid growth of long-read sequencing, tools like FLAIR will be useful in identifying key disease-associated variants that may serve as biomarkers of potential prognostic or therapeutic relevance.

1.12 Methods

Data generation and handling

Peripheral blood mononuclear cells were obtained from patients with CLL and from healthy adult volunteers, enrolled on sample collection protocols at Dana-Farber Cancer Institute, approved by and conducted in accordance with the principles of the Declaration of Helsinki and with the approval of the Institutional Review Board (IRB) of Dana-Farber Cancer Institute. Samples were cryopreserved in 10% DMSO until the time of RNA extraction. RNA was extracted from tumor samples using methods previously described (L. Wang et al. 2016) The sample IDs of the CLL *SF3B1*^{WT} samples are CW67 (WT 1), CW95 (WT 3), and JGG0035 (WT 2) and the IDs of the *SF3B1*^{K700E} samples are DFCI-5067 (MT 1), CLL043/CW109 (MT 2), and CLL032/CW84 (MT 3) from Wang et al. (L. Wang et al. 2016) JGG035 is the only sample not included in that study. All samples had RIN scores above 7. The extracted RNA was reverse transcribed using the SmartSeq protocol (Picelli et al. 2013) and cDNA was PCR-amplified as

described in Byrne et al. (Byrne et al. 2017). 15 cycles of PCR were performed. Prior to library preparation, the concentration of the cDNA for the samples ranged from 1.26-10.7 ng/ul. Oxford MinION 2D amplicon libraries were generated according to the Nanopore community protocol using library preparation kit SQK-LSK208 and sequenced on R9 flowcells. Basecalling was performed with albacore v1.1.0 2D basecalling using the --flow cell FLO-MIN107 and --kit SQK-LSK208 options. The same cDNA preparation protocol was used for PromethION sequencing. Library preparation for 1D sequencing was performed following Oxford Nanopore's protocol, with the exception of the last bead clean up using a 0.8x bead ratio. The PromethION libraries were prepared and sequenced in one batch of 3 and one batch of 6, with at least 1 sample of each condition in each batch. Basecalling of 1D PromethION reads was done with guppy v2.3.5 with the default options, and only reads that were designated "pass" reads in the summary file were used for subsequent analyses. We identified reads with adapter sequences on both ends following the approach employed in the MandalorION pipeline (Byrne et al. 2017): (1) adapters are aligned to all the reads using blat (Kent 2002), (2) if there are at least 10 bases at the left and right ends of the reads that match the adapter sequence then the read is considered to have adapters on both ends. We found that only a fraction of our reads that were called as "pass" reads contained the adapter sequences on both ends (~35-55%). In the interest of being able to use more of our data, we did not remove these reads from the analyses.

Nanopore sequencing statistics

The reads for each sample were aligned with minimap2 v2.7-r654 (Li 2016) to the GENCODE v24 transcriptome and the read-isoform assignments were determined using the primary alignments. Following read-isoform assignment, the percentage of full-length reads was calculated as the number of reads covering 80% of nucleotides for the transcript they were assigned to divided by the total number of reads that aligned. The number of genes observed was computed by counting the genes represented by all the isoforms. Genes with multiple isoforms identified were considered alternatively spliced.

Spliced alignment and read correction

Reads were aligned to the hg38 genome downloaded from UCSC (http://hgdownload.cse.ucsc.edu/goldenPath/hg38/bigZips/) using minimap2 v2.7-r654 (Li 2016) in spliced alignment mode with the command `minimap2 -ax splice`. GMAP 2017-10-30 (Wu and Watanabe 2005) was used for comparison against minimap2. Indels were removed from the read alignments. FLAIR *correct* (v1.4) was used to correct the splice site boundaries of reads. All splice sites were assessed for validity by checking for support in GENCODE v24 comprehensive annotations or short reads. Splice junctions were extracted from matched short-read data and only the junctions supported by 3 uniquely mapping short reads were considered valid. Incorrect splice sites were replaced with the nearest valid splice site within a 10-nt window. The set of corrected reads consists of reads that contain only valid splice sites.

Isoform identification methods

For running FLAIR on the PromethION CLL/B cell data, the following FLAIR *collapse* algorithm was followed: to assemble the first-pass assembly, transcription start sites and transcription end sites are determined by the density of the read start and end coordinates. We compared 100 nt windows of end sites and picked the most frequently represented site in each window (-n best_only). The final nanopore-specific reference isoform assembly is made by aligning raw reads to the first-pass assembly transcript sequence using minimap2, keeping only the first-pass isoforms with a minimum number of 3 supporting reads with MAPQ>=1. All pass reads, including reads that did not contain sequenced adapters on both ends, were used when running FLAIR as FLAIR is equipped to deal with truncated reads; information can be gleaned from truncated reads of sufficient length to be assigned to an isoform and the reads that are too short for a unique assignment are excluded from the isoform quantification.

Saturation analysis

We performed the saturation analysis on the 3 runs with the most coverage in each sample (WT 3, MT 2, B Cell 1). The total reads from each run was used, in addition to subsampled sets of reads. Reads were subsampled in increments of 10 million by random selection using python random.sample(). We used FLAIR to identify isoforms within each subset of reads using the `best_only` parameter to obtain only one transcription start and end site per splice junction chain.

Isoform quantification and fold-change calculation

Isoforms were quantified using FLAIR *quantify*, only counting the alignments with quality scores of 1 or greater. Isoform counts within each sample were normalized by dividing each count by the upper quartile (75th percentile) of the read counts of protein-coding genes. Only genes labeled as protein coding in GENCODE v24 annotation were considered protein-coding.

Alternative splicing event calling and statistical testing

Custom scripts were written for FLAIR to identify alternative acceptor, alternative donor, cassette exon skipping, and intron retention events (FLAIR *diffSplice*). Alternative 3'SS were grouped by the 5'SS they were observed with and had to be present in overlapping exons and vice versa for alternative 5'SS calling. Alternative 3' and 5' splice sites that were within 10 bp of each other were exempt from statistical tests. For the analysis of the pilot data containing one wild type and one mutant *SF3B1*, a Fisher's exact test was used to determine the significance of alternative splicing events. For analysis of the PromethION data with replicates, we used DRIMSeq (Nowicka and Robinson 2016). DRIMSeq statistical testing accounted for sequencing and RNA batch according to batch numbers for each sample. The expression filters used for DRIMSeq were as follows: a minimum of 4 of the 6 samples should cover either the inclusion or exclusion event with minimum coverage of 25 reads. Of the 4 samples with sufficient coverage, 2 were required to be from either the CLL $SF3B1^{WT}$ or $SF3B1^{K700E}$ condition.

A pseudocount of 1 was used to prevent events with 1-2 dropout samples from being excluded from testing. For differential isoform usage testing (FLAIR *diffExp*), isoforms were grouped by gene and only genes with at least 25 reads in 4 of the 6 samples were tested. We did not distinguish between intron retentions due to incomplete transcript processing and intron retentions deliberately retained due to sample genotype.

Intron retention and productivity analysis

Fold-change was calculated using the median upper-quartile-normalized isoform count for each condition and dividing the mutant expression by the wild type expression. Only the transcripts with a median of 10 or more in one of the conditions were plotted. Intron retentions were defined as any intron that is completely spanned by another isoform's exon. For identification of NMD-sensitive transcripts, we used annotated start codons from GENCODE v24 and translated the full-length assembled isoforms. Isoforms with a PTC were called unproductive, and isoforms without PTCs were called productive. A PTC was defined as a stop codon detected before 55 nucleotides or more upstream of the last splice junction (Rivas et al. 2015). If a transcript overlapped more than one annotated start codon, the productivity was assessed by using (1) the 5' most start codon or (2) the start codon yielding the longest transcript; if both strategies yielded different productivity results, then the isoform was excluded from analysis.

GO analysis

GO analysis was performed with the R package goseq v1.32.0 (Young et al. 2012), setting the parameter method=hypergeometric to remove the correction for gene length bias that affects short-read data. GO terms with only one term in the category were removed from further analysis.

Chapter 2: Direct RNA sequencing for the characterization of poly(A) RNAs

1.1 Abstract

High-throughput complementary DNA sequencing technologies have advanced our understanding of transcriptome complexity and regulation. However, these methods lose information contained in biological RNA because the copied reads are often short and modifications are not retained. We address these limitations using a native poly(A) RNA sequencing strategy developed by Oxford Nanopore Technologies. Our study generated 9.9 million aligned sequence reads for the human cell line GM12878, using thirty MinION flow cells at six institutions. These native RNA reads had a median length of 771 bases, and a maximum aligned length of over 21,000 bases. We combined these long nanopore reads with higher accuracy short-reads and annotated GM12878 promoter regions to identify 33,984 plausible RNA isoforms, updating FLAIR in the process to adequately deal with the challenges of direct RNA data. We open up strategies for assessing 3′ poly(A) tail length, base modifications and transcript haplotypes, although the focus in this dissertation will be on haplotype-specific transcripts.

2.2 Introduction

Sequencing by synthesis (SBS) strategies have dominated RNA sequencing since the early 1990s (Adams et al. 1991). They involve generation of cDNA templates by reverse transcription (Temin and Mizutani 1970; Baltimore 1970) coupled with PCR amplification (Saiki et al. 1988). Nanopore RNA strand sequencing, or direct RNA sequencing, has emerged as an alternative single molecule strategy (Garalde et al. 2018; Jenjaroenpun et al. 2018; A. M. Smith et al. 2019). It differs from SBS-based platforms in that native RNA nucleotides, rather than copied DNA nucleotides, are identified as they thread through and touch a nanoscale sensor. Nanopore RNA strand sequencing shares the core features of nanopore DNA sequencing, *i.e.*

a processive helicase motor regulates movement of a bound polynucleotide driven through a protein pore by an applied voltage. As the polynucleotide advances through the nanopore in single nucleotide steps, ionic current impedance reports on the structure and dynamics of nucleotides in or proximal to the channel as a function of time. This continuous ionic current series is converted into nucleotide sequence using an ONT neural network algorithm trained with known RNA molecules.

Here we describe sequencing and analysis of a human poly(A) transcriptome from the GM12878 cell line using the Oxford Nanopore (ONT) platform. We demonstrate that long native RNA reads allow for the discovery and characterization of polyA RNA molecules that are difficult to observe using short read cDNA methods (Steijger et al. 2013; Venturini et al. 2018). Data and resources are posted online at: (https://github.com/nanopore-wgs-consortium/NA12878/blob/master/RNA.md).

2.3 Sequencing summary

Six laboratories each performed five nanopore sequencing runs. These thirty runs produced 13.0 million poly(A) RNA strand reads, of which 10.3 million passed quality filters (PHRED>7). Throughput varied between 50K and 831K pass reads per flow cell, with an N50 length of 1,334 bases, and a median of 771 bases. Of these, 9.9 million aligned using minimap2 (H. Li 2018) to the GRCh38 human genome reference. The 360,000 unaligned pass reads had a median read length of 211 bases.

2.4 FLAIR for improved isoform detection in direct RNA data

Long nanopore reads could improve resolution of RNA exon-exon connectivity, allowing for discovery of unannotated RNA isoforms. However, these reads averaged 14% per-read basecall errors, confounding precise determination of splice sites. Also, biological RNA processing and *in vitro* 5′-end truncations can make it difficult to define transcription start sites (TSS).

To overcome these limitations we employed FLAIR (Tang et al. 2020) (Full-Length Alternative Isoform Analysis of RNA). We first replaced any nanopore-based splice sites bearing apparent sequencing errors with splice sites supported by GENCODE v27 annotations or by Illumina GM12878 cDNA data (Fig 2.1) (Tilgner et al. 2014; Cho et al. 2014). Second, to overcome TSS uncertainty caused by truncated RNA reads, we considered only reads with 5′ ends proximal to promoter regions as defined by ENCODE promoter chromatin states for GM12878 (Bernstein et al. 2005; Ernst and Kellis 2010; Ernst et al. 2011). Third, we used FLAIR to group reads into isoforms according to unique chains of splice junctions.

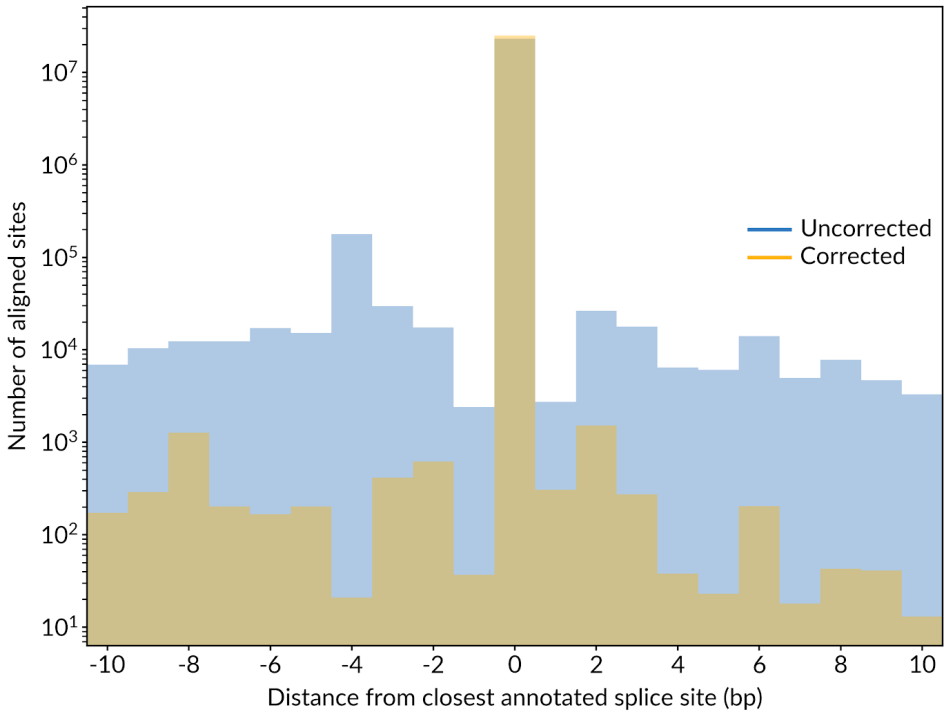

Fig 2.1 Correcting minimap2 genomic read alignments improves splice site accuracy. *Using FLAIR-correct, misaligned splice sites were corrected to splice sites supported by short-read sequencing. The x-axis is the distance from the aligned splice site to the closest annotated splice site in GENCODE v27.*

The y-axis is the number of aligned sites (log-scale) with raw alignment distance counts in blue and corrected counts in yellow.

We compiled two FLAIR isoform sets using different supporting read criteria (Methods):

i) A FLAIR-sensitive set that included isoforms with three or more uniquely mapped reads. This large set could be useful for isoform discovery, at the risk of false positives.

ii) A FLAIR-stringent set that was compiled by filtering set (i) for isoforms having three or more supporting reads that spanned ≥80% of the isoform with ≥25nt coverage into the first and last exon (Fig 2.2).

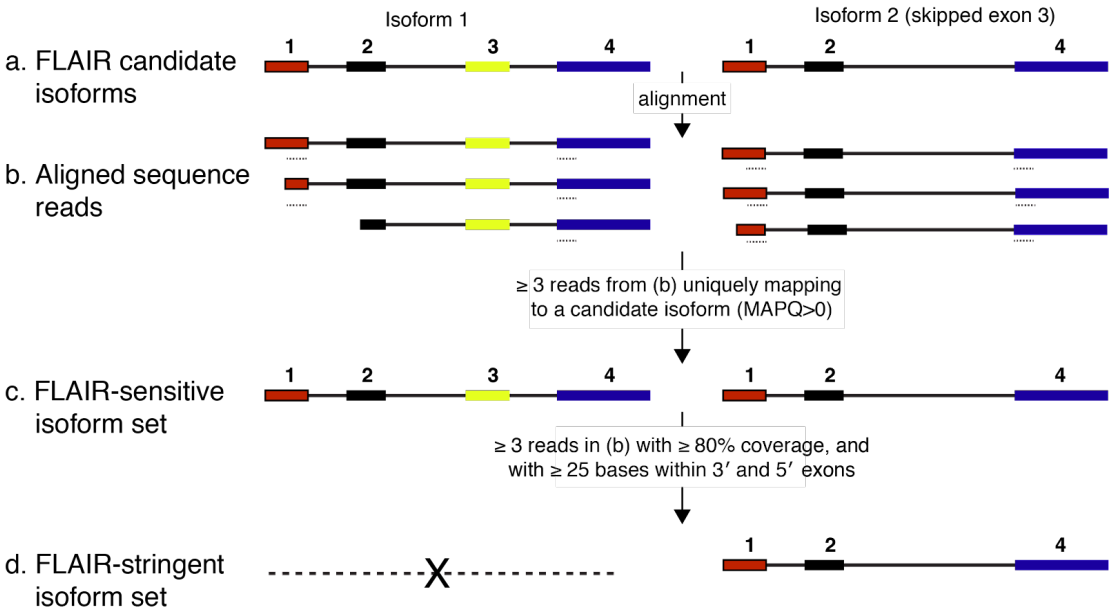

Figure 2.2 Criteria for the FLAIR-sensitive and FLAIR-stringent isoform sets. a *Two candidate isoforms assembled using FLAIR. Each block represents either a complete or a partial exon (numbers 1-4).* ***b*** *Reads that align to a candidate isoform. Light gray bars represent 25 nt coverage into first and last exons.* ***c*** *FLAIR-sensitive isoform set that passed criteria shown at arrow.* ***d*** *FLAIR-stringent isoform set that passed criteria shown at arrow. Isoform 1 failed FLAIR-stringent isoform test (X); isoform 2 passed FLAIR-stringent isoform test.*

We screened for unannotated isoforms within the FLAIR-stringent dataset. Of the 33,984 isoforms representing 10,793 genes, 52.6% had a splice junction chain that was unannotated in GENCODE (13.0% of total assigned reads). Figure 2.3a shows an example set of lncRNA isoforms arising from an unannotated transcription start site with multiple splice variants. We observed that non-coding genes had more complex splicing patterns per gene than did coding genes (Figure 2.3b), in agreement with prior studies demonstrating increased alternative splicing in non-coding exons (Deveson et al. 2018; Gonzàlez-Porta et al. 2013).

As a conservative alternative to FLAIR, we compiled two GENCODE-based isoform sets:

i) A GENCODE-sensitive set that included isoforms with one or more reads that mapped uniquely to GENCODE v27. We implemented a lower coverage threshold than we did for FLAIR because GENCODE is curated.

ii) A GENCODE-stringent set that was compiled by filtering set (iii) for isoforms having one or more supporting reads that spanned ≥80% of the isoform with ≥25nt coverage into the first and last exon.

To estimate the sequencing depth required to completely characterize the GM12878 transcriptome, we plotted the number of isoforms detected in the GENCODE-sensitive and FLAIR-stringent isoform sets versus the number of subsampled reads in 10% increments. We then fitted a hyperbolic function to the data (Figure 2.3c**).** It is evident that the curves did not saturate and that additional reads would be required to capture a complete GM12878 transcriptome.

2.5 Assignment of transcripts to parental alleles

Allele-specific expression (ASE) is the preferential transcription of RNA from the paternal or maternal copy of a gene. Although the importance of this phenomenon has been characterized (Baralle and Giudice 2017), the consequences are not fully understood. This is

partly due to technical limitations of haplotype identification using short read sequencing technologies.

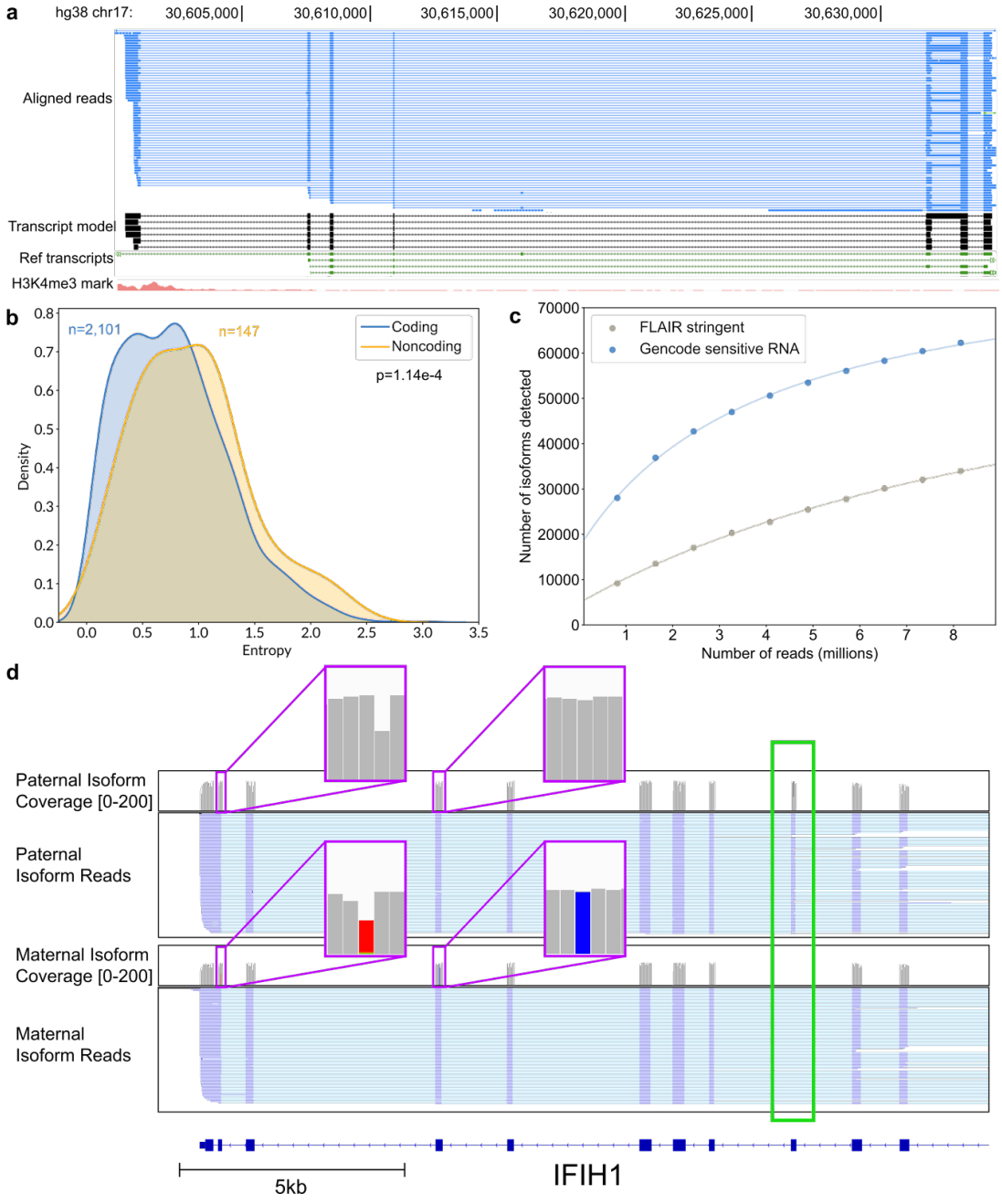

Figure 2.3 Isoform-level analysis of GM12878 native poly(A) RNA sequence reads. a *Genome browser view of unannotated isoforms that aligned to SMURF2P1-LRRC37BP1. The tracks are: a subset*

of the aligned native RNA reads (blue); the FLAIR-defined isoform models (black); SMURF2P1-LRRC37BP1 annotated isoforms from GENCODE v27 comprehensive set (green); transcription regulatory histone methylation marks (red). ***b*** *Shannon entropy of isoform expression for coding versus noncoding genes detected by FLAIR. The p-value was calculated using a Mann-Whitney U test.* ***c*** *Saturation plot showing the number of isoforms discovered (y-axis) versus the number of native RNA reads (x-axis).* ***d*** *IGV view of allele-specific isoforms for IFIH1. Purple boxes (insets) indicate the location of SNPs used to assign allele specificity (gray reference; red and blue SNPs). The alternatively spliced exon is indicated by a green box.*

We reasoned that the long nanopore RNA reads would be easier to assign to the parental allele of origin due to the greater chance of encountering a heterozygous SNP. Reads with at least two heterozygous SNPs were assigned to the parental allele of origin using HapCUT2 (Edge, Bafna, and Bansal 2017). To discover the most possible genes, we used the FLAIR-sensitive dataset. In it, we found 3,751 genes with at least 10 haplotype informative reads. Among autosomal genes, 228 (6.1%) showed significant ASE (binomial test, $p<0.001$), and among X-chromosome genes, 23 (95.7%) showed significant ASE (binomial test, $p<0.001$). X-chromosome expression was biased, with 22/23 allele-specific X-linked genes originating from the maternal allele, consistent with previous results for this cell line (Rozowsky et al. 2011). The sole paternally expressed X-linked locus encoded the lncRNA XIST, which is transcribed from the inactive X-chromosome and recruits epigenetic silencing machinery for X-inactivation in females (Brown et al. 1991). The remaining genes were expressed equally from both parental alleles.

We combined these allele-specific reads with isoforms from the FLAIR-sensitive set to mine for allele-specificity (Methods). We identified 5 genes with one isoform expressed from one allele and another isoform expressed from the other allele (binomial test, $P<0.001$). One of these genes, *IFIH1*, had a paternal isoform with exon 8 retained, while the maternal isoform did not retain exon 8 (Figure 2.3d). We note that the closest SNV used in allele-assignment

was 886 nt away from the alternative splicing event in this transcript. This would be undetectable using short read sequencing.

2.6 3′ poly(A) analysis

Transcript poly(A) tails are thought to play a role in post-transcriptional regulation, including mRNA stability and translational efficiency (Eckmann, Rammelt, and Wahle 2011). However, these homopolymers can be several hundred nucleotides long making them difficult to measure using short-read SBS data (Subtelny et al. 2014; Chang et al. 2014). Workman et al. measured poly(A) tail lengths directly using a low variance ionic current signal associated with the 3′ end of each poly(A) strand. Nanopolish-polya (https://github.com/jts/nanopolish) is a computational method to segment this signal and estimate how many ionic current samples were drawn from the poly(A) tail region. By correcting for the rate at which the RNA molecule passes through the pore, nanopolish-polya estimates the length of the poly(A) tail.

We applied this poly(A) length estimator to the complete GM12878 native poly(A) RNA sequence dataset. Nuclear transcripts showed a broader length distribution, with a peak at 58nt, a mean of 112nt, and a large number of poly (A) tails greater than 200nt. We analyzed genes in the GENCODE-sensitive dataset, exploring the relationship between poly(A) tail length and RNA intron-retention. We classified each isoform in GENCODE-sensitive as either protein-coding or intron-retaining with FLAIR. The subset of transcripts with retained introns tended to have longer poly(A) tails (median 232nt) than did transcripts without introns (median 91nt) (t-test p-value < 2.2e-16).

2.7 Noncanonical base detection

Nanopore sequencing has been used to identify base modifications in DNA (Simpson et al. 2017; Rand et al. 2017) and RNA (Garalde et al. 2018; A. M. Smith et al. 2019). One example of a modified base arises from adenosine-to-inosine RNA editing (Licht et al. 2016), which plays a role in splicing and regulating innate immunity (Nishikura 2010; Tajaddod,

Jantsch, and Licht 2016). Illumina sequencing detects A-to-I editing as an A-to-G nucleotide variant in cDNA sequences.

Previous nanopore experiments documented the presence of systematic base miscalls in regions of *E. coli* 16S rRNA bearing modified RNA bases (A. M. Smith et al. 2019). We found systematic base miscalls at putative inosine bearing positions in the GM12878 aryl hydrocarbon receptor (*AHR*) data. To cross-validate, we compared our cDNA sequence data relative to the GM12878 reference and found that putative inosines were detected as an A-to-G base change as expected (*i.e.* a single inosine for the CUACU 5-mer, and multiple inosines for the AAAAA 5-mer).

The ionic current distribution for the putative single inosine 5-mer (CUACU) was modestly different from the canonical 5-mer. The ionic current distribution for the inosine containing AAAAA 5-mer was more complex, possibly reflecting the presence of multiple inosines.

2.8 Discussion

Nanopore RNA sequencing has two useful features: 1) The sequence composition of each strand is read as it existed in the cell. This permits direct detection of post-transcriptional modifications including nucleotide alterations and polyadenylation; 2) reads can be continuous over many thousands of nucleotides providing splice-variant and haplotype phasing. Although each of these features is useful in itself, the combination is unique and likely to provide new insights into RNA biology. The two principal drawbacks of the present ONT nanopore RNA sequencing platform is the relatively high error rate (compared to Illumina cDNA sequencing), and uncertainty about the 5' end of the transcript.

We were concerned about read fragmentation due to RNA degradation during sequencing. However, we found minimal (~5%) reduction in the full-length fraction of a 1.6 kb mRNA (*MT-CO1*) over 36 hours. Preliminary analysis indicated that read truncations were

more often caused by electronic signal noise due to current spikes of unknown origin. We showed that meaningful biological signals can be recovered from bulk Fast5 files around these truncations, suggesting that future improvements to the MinKNOW read segmentation pipeline are needed.

When combined with more accurate short Illumina reads, long nanopore reads allowed for end-to-end documentation of RNA transcripts bearing numerous splice junctions, which would not be possible using either platform alone. We documented a high proportion (52.6%) of unannotated isoforms, similar to other long-read transcriptome sequencing studies (e.g., 35.6% and 49%) (Tardaguila et al. 2018; Anvar et al. 2018). While many of these unannotated isoforms are low abundance and their protein coding potential unknown, it is important to catalog them because subtle splicing changes can impact function (L. Wang et al. 2016; Bradley et al. 2012). We also note that the number of detected isoforms did not saturate using the nanopore poly(A) RNA dataset, indicating that greater sequence depth will be necessary to give a comprehensive picture of the GM12878 poly(A) transcriptome.

A variety of techniques have been used to examine allele-specific expression (Rozowsky et al. 2011; Tilgner et al. 2014). However, identification of ASE is limited using short read platforms because heterozygous variants are rare within any given window of a few hundred nucleotides. Nanopore sequencing has the advantage of long reads, albeit limited by errors. We have shown that nanopore sequencing enables allele-specific isoform studies. With further work on haplotype-calling in nanopore data, we expect to be able to detect haplotype-specific transcripts, particularly for cases where the splicing variation does not have a heterozygous variant within range of conventional short-read sequencing.

Polyadenylation of RNA 3′ ends regulates RNA stability and translation efficiency by modulating RNA-protein binding and RNA structure (Eckmann, Rammelt, and Wahle 2011). However, transcriptome-wide poly(A) analysis has been difficult due to basecalling and dephasing errors (Chang et al. 2014). Recently implemented modifications to the Illumina

strategy address these limitations (Chang et al. 2014; Subtelny et al. 2014); but can not resolve distal relationships, such as between splicing and poly(A) length. Nanopore poly(A) tail length estimation using nanopolish-polya (Workman et al. 2018) offers the advantages of both direct length assessment and maintenance of information about isoform and modification status per transcript. Our preliminary studies revealed differences in poly(A) length distribution between mitochondrial and nuclear genes, between different nuclear genes, and between different isoforms of the same gene (Workman et al. 2018). We note an increase in poly(A) tail length for some intron-retaining isoforms. This is consistent with previous work showing that hyper-adenylation targets intron-retaining nuclear transcripts for degradation through recognition by a poly(A)-binding protein (PABPN1) (Bresson et al. 2015). Additionally, deadenylation of cytoplasmic transcripts is a core part of the RNA degradation pathway (Yi et al. 2018), suggesting that time course experiments investigating RNA decay kinetics (Parker and Song 2004) could be possible with this technology.

Although other methods exist for high throughput analysis of RNA modifications (X. Li, Xiong, and Yi 2016), they often require enrichment which limits quantification, and they are usually short-read based. The latter precludes analysis of long-distance interactions between modifications, and between modifications and other RNA features such as splicing and poly(A) tail length. The capacity to detect these long-range interactions is likely to be important given recent work suggesting links between RNA modifications, splicing regulation, and RNA transport and lifetime (Roundtree et al. 2017; Lee, Kim, and Kim 2014). We argue that nanopore native RNA sequencing could deliver this long-range information for entire transcriptomes. However, this will require algorithms trained on large, cross-validated datasets as has been accomplished for cytosine and adenine methylation in genomic DNA (Simpson et al. 2017; Rand et al. 2017).

2.9 Methods

Isoform detection and characterization

To define isoforms from the sets of native RNA and cDNA reads, we used FLAIR v1.4, a version of FLAIR(Tang et al. 2018) with additional considerations for native RNA nanopore data. For our analysis, we first removed reads generated by lab 6, because a disproportionate number of those molecules appeared to be truncated prior to addition to the nanopore flow cell. We also removed 71,276 aligned reads with deletions greater than 100 bases caused by minimap2 version 2.1. We then selected reads that had TSSs within promoter regions that were computationally derived from ENCODE ChIP-Seq data(Ernst and Kellis 2010; Ernst et al. 2011). Using FLAIR-correct, we corrected primary genomic alignments for pass reads based on splice junction evidence from GENCODE v27 annotations and Illumina short-read sequencing of GM12878. This step also removes reads containing non-canonical splice junctions not present in the annotation or short-read data. The filtered and corrected reads were then processed by FLAIR-collapse which generates a first-pass isoform set by grouping reads on their splice junctions chains. Next, pass reads were realigned to the first-pass isoform set, retaining alignments with MAPQ>0. Isoforms with fewer than 3 supporting reads or those which were subsets of a longer isoform were filtered out to compile the FLAIR-sensitive isoform set. A FLAIR-stringent isoform set was also compiled by filtering the FLAIR-sensitive set for isoforms which had 3 supporting reads that spanned ≥80% of the isoform and a minimum of 25nt into the first and last exons. Unannotated isoforms were defined as those with a unique splice junction chain not found in GENCODE v27. Isoforms were considered intron-retaining if they contained an exon which completely spanned another isoform's splice junction. Isoforms with unannotated exons were defined as those with at least one exon that did not overlap any existing annotated exons in GENCODE v27. Genes that did not contain an annotated start codon were considered non-coding genes.

Defining promoter regions in GM12878 for isoform filtering

Promoter chromatin states for GM12878 were downloaded from the UCSC Genome Browser in BED format from the hg18 genome reference. Chromatin states were derived from an HMM based on ENCODE ChIP-Seq data of nine factors (Ernst and Kellis 2010; Ernst et al. 2011). The liftover tool (Hinrichs et al. 2006) was used to convert hg18 coordinates to hg38. The active, weak, and poised promoter states were used.

Calculating isoform entropy of genes

Productivity was assessed according to the NMD rule followed in the CLL study where if a premature termination is located 55 nt or more upstream of the last exon-exon junction, the transcript is considered unproductive (Rivas et al. 2015). Genes that did not contain an annotated start codon were considered noncoding genes. Only genes with at least 50 reads as well as more than two isoforms were considered for the entropy analysis.

Chapter 3: Knockdown of ADAR to interrogate A-to-I editing in lung adenocarcinoma progression

Abstract

RNA-Seq has brought forth significant discoveries regarding aberrations in RNA processing, implicating these RNA variants in a variety of diseases. In particular, aberrant splicing and single nucleotide variants in RNA have been demonstrated to alter transcript stability, localization, and function. Despite the functional importance of studying splicing and SNVs, short read RNA-Seq has limited the community's ability to interrogate both forms of RNA variation simultaneously. Thus, we have employed long-read technology to obtain full-length transcript sequences, elucidating cis-effects of variants on splicing changes at a single molecule level. We have developed a computational workflow that augments FLAIR, a tool that calls isoform models expressed in long-read data, to integrate RNA variant calls with the associated isoforms that bear them. Applying this pipeline to an F1 hybrid mouse embryonic stem cell line (castaneus x S129/SvJae) sequenced from the Long-read RNA-Seq Genome Annotation Assessment Project, we are able to identify allele-specific isoform expression connected to each parent. Additionally, we have generated nanopore data of H1975 lung adenocarcinoma cells with and without knockdown of ADAR. Upregulation of ADAR, an enzyme which mediates adenosine-to-inosine editing, has been previously linked to an increase in the invasiveness of lung ADC cells and has been linked to the regulation of splicing. We applied our workflow to identify key inosine-isoform associations to help clarify the prominence of ADAR in tumorigenesis. Ultimately, we find that a long-read approach provides valuable insight toward characterizing the relationship between RNA variants and splicing patterns.

Introduction

Adenosine-to-inosine (A-to-I) editing is one of the most common forms of RNA editing in organisms with a developed central nervous system (Athanasiadis, Rich, and Maas 2004; Levanon et al. 2004; Nishikura 2010; Kiran et al. 2013; Bazak et al. 2014). As inosines are recognized by cellular machinery as a guanosine, one potential downstream effect of A-to-I editing is the alteration of coding sequence. There are numerous cases of A-to-I recoding identified as essential for normal brain function (Sommer et al. 1991; Burnashev et al. 1992; Bajad et al. 2017) and yet other cases where recoding worsens disease prognosis (Han et al. 2015; Amin et al. 2017; Lazzari et al. 2017). In addition to recoding potential, inosines can affect RNA splicing in a *cis*-regulatory manner through the disruption of splice sites or splicing regulatory elements, leading to the creation of alternatively spliced mRNAs (Rueter, Dawson, and Emeson 1999; Hsiao et al. 2018; S. J. Tang et al. 2020). Considering that 95-100% of multi-exon genes are alternatively spliced (Pan et al. 2008), the effects of ADARs on coding changes, regulatory elements, and alternative splicing require further study to elucidate.

The expression of two ADAR family proteins, ADAR1 and ADAR2, is ubiquitous and the edits are widespread in mRNAs (C. X. Chen et al. 2000). Aberrant ADAR activity has been linked to many diseases: 1) related to amyotrophic lateral sclerosis, a decrease in the efficiency of A-to-I editing detrimentally increases Ca^{2+} permeability in neurons (Sommer et al. 1991; Burnashev et al. 1992; Kawahara et al. 2003); 2) mutations in *ADAR1* have been shown to cause Aicardi-Goutières syndrome (Livingston et al. 2014; Rice et al. 2012); 3) in breast cancer, A-to-I editing of *Gabra3* mRNA suppresses an invasive phenotype (Gumireddy et al. 2016); 4) in hepatocellular carcinoma, ADAR recoding stabilizes the AZIN1 protein leading to increased cell proliferation (L. Chen et al. 2013); and 5) in diseases of the lung and blood diseases, ADAR overexpression is associated with increased malignancy (Amin et al. 2017; Lazzari et al. 2017). Additionally, in H1975 lung adenocarcinoma (ADC) cell lines, ADAR is not only upregulated but also has been shown to bind to and edit focal adhesion kinase (FAK), increasing both FAK expression and mesenchymal properties of the cells (Amin et al. 2017). The connection of

ADARs with diseases, in particular lung adenocarcinoma, in addition to the influence that ADARs have on the transcriptome underscores the importance of characterizing the complete RNA sequences that bear inosine edits.

Despite appreciable efforts to map A-to-I editing sites (Ramaswami and Li 2014; Kiran et al. 2013), there is an absence of studies examining the full transcriptional context of inosines. Previous efforts to document A-to-I editing using short-read sequencing report only the genomic position of edited sites (Ramaswami and Li 2014; Kiran et al. 2013; Picardi et al. 2015). To investigate the transcriptome-wide impact of ADAR in lung ADC, we performed nanopore long-read cDNA sequencing of lung ADC cells with ADAR knockdown. The relatively high error rate of nanopore sequencing hinders work that relies on high sequence accuracy (Workman et al. 2018); we overcome this setback by using the Rolling Circle Amplification to Concatemeric Consensus (R2C2) nanopore cDNA sequencing method (Volden et al. 2018). R2C2 greatly lowers the error rate of nanopore cDNA sequencing through the increase of single molecule coverage, yielding a median 98.7% base accuracy (Byrne et al. 2019). Accurate, long reads allow us to resolve full-length transcripts and RNA editing, equipping us to better understand the role of ADAR editing in the cancer transcriptome.

Nanopore sequencing operates on the sensing of changes in current as genetic material passes through a nanopore (Deamer, Akeson, and Branton 2016). The current signal associated with modified RNA bases can cause modifications to be misbasecalled as the incorrect canonical base, thus appearing as a mismatch to the reference genome once the sequence is aligned (A. M. Smith et al. 2019b). In direct nanopore reads, there is ambiguity as to whether mismatches to the reference correspond to somatic or germline variants, RNA edits, or RNA modifications. While R2C2 is unable to preserve RNA modifications, we have devised a tool to phase and associate mismatches to isoform models in long reads, agnostic to the kind of alteration that generated the mismatch. We refer to these mismatch-aware isoforms generally as haplotype-specific transcripts (HSTs). There is a lack of available computational

software for identifying HSTs, necessitating the development of a tool to jointly identify isoform structure and inosine positions in nanopore data. We built upon the isoform detection tool FLAIR, which is one amongst many tools (Stringtie2, FLAMES, TALON, MandalorION) developed for this purpose. FLAIR was initially developed to identify subtle splice site changes in long reads with higher error rates and increased truncation. Our variant-aware FLAIR, called FLAIR2, differs from the LORALS (Glinos et al. 2021) and IDP-ASE (Deonovic et al. 2017) tools in that FLAIR2 can incorporate mismatches in transcript models for an arbitrary number of haplotypes.

Here, we sequenced three replicates with ADAR1 knocked down and three replicates receiving a negative control using Illumina RNA-Seq as well as R2C2 nanopore sequencing. With the development of the necessary computational framework for full-length isoform and RNA editing analyses, we reveal new insights into longer-range A-to-I edits and demonstrate the power of nanopore sequencing as a tool for the transcriptome-wide identification of inosines.

FLAIR2 is a variant-aware isoform detection pipeline

In an effort to build user-friendly computational workflows for nanopore data, we previously had developed a computational tool called Full-Length Alternative Isoform analysis of RNA (FLAIR). FLAIR calls isoform structures and performs various isoform-level analyses of nanopore cDNA (A. D. Tang et al. 2018) and nvRNA (Workman et al. 2018; Soneson et al., n.d.) data. We have designed the FLAIR workflow to account for the increased error rate of long reads. Previous work with FLAIR emphasized the discovery of isoform models and their comparison between sample conditions. We have adjusted FLAIR to incorporate phased variant calls to investigate haplotype-specific transcript expression in nanopore data. We have also improved FLAIR's performance on SIRV isoform identification precision and sensitivity.

The modified FLAIR workflow now begins with an alignment of all reads to the annotated transcriptome. The addition of this ungapped alignment step is for the cases where

genomic alignment of the long, spliced read is difficult for aligners, such as microexons (B. Liu et al. 2019). Reads are assigned to an annotated transcript if they have high sequence identity with the transcript, with an emphasis of accuracy proximal to splice sites (see Methods). The annotated transcripts that have sufficient long read support are included as part of the set of FLAIR isoforms. The remaining reads that are not able to be assigned to an annotated transcript are then used to detect novel transcript models (see Methods). The final, sample-specific isoform assembly includes the supported, annotated isoform models combined with the novel models. FLAIR is also capable of downstream analyses such as isoform quantification and differential expression tests of nanopore data. FLAIR is on GitHub at https://github.com/BrooksLabUCSC/flair.

Assessing FLAIR2 for haplotype-specific transcript detection

We compared the performance of FLAIR2's updated isoform detection method with Stringtie2 and FLAMES on SIRVs sequenced with nanopore R2C2 sequencing (see Methods). Transcript detection with FLAIR2 is more precise than other tools (Table 1), indicating fewer false positive transcripts detected with FLAIR2. The sensitivity between the tools are comparable. We also investigated the transcript-level precision and sensitivity using nanopore 1D cDNA SIRV sequences (Table 2), where FLAIR2 again performed best comparatively in precision and performed similarly to other tools in terms of sensitivity. On these SIRVs, FLAIR2 demonstrated marked improvement over the previously published FLAIR, which focused more on base-level sensitivity and precision.

	Transcript-level sensitivity	Transcript-level precision
FLAIR2	77.8	95.5
Stringtie2	72.8	90.8
FLAMES	81.5	94.3

Table 1 Performance of transcript detection tools on R2C2 nanopore SIRVs.

	Transcript-level sensitivity	Transcript-level precision
FLAIR2	63.8	89.8
Stringtie2	76.8	81.5
FLAMES	66.7	78.0
FLAIR	65.1	51.9

Table 3 Performance of transcript detection tools on 1D nanopore SIRVs.

We tested both longshot (Edge and Bansal 2019) and PEPPER-Margin-DeepVariant (Shafin et al. 2021) to call variants in long-read data. Both variant callers can perform diploid variant calling and phasing. FLAIR has two modalities for variant-aware transcript detection. One, FLAIR can incorporate phased variants, such as those provided by longshot, which have information pertaining to the phase set a read is assigned to. Two, as we anticipated working with RNA edits and potential cancer-related aneuploidies that may result in more than two apparent haplotypes, FLAIR takes a more simplistic approach to phasing alignment mismatches that is agnostic to ploidy: 1) given variant calls, FLAIR tabulates the most frequent combinations of variants present in each isoform from the supporting read sequences; 2) from the isoform-defining collapse step, FLAIR generates a set of reads assigned to each isoform; so 3) isoforms that have sufficient read support for a collection of mismatches are determined (Fig 3.1a). This accommodates for multiple haplotypes within a gene and within a transcript model.

We tested the FLAIR2 isoform discovery pipeline on Castaneus x Mouse 129 hybrid mouse embryonic stem cells where we expect evidence of HSTs partitioned by parental haplotypes. Integrating longshot's phased diploid variant calls, we identified 1,017 genes that contained HSTs with FLAIR2. One example is shown in Fig 3.1b and e, in which the non-reference haplotype that longshot reports, which corresponds to the castaneus parent haplotype, is biased toward the expression of isoforms with a proximal 5' splice site. With this, we determined that FLAIR2 can be used to detect HSTs successfully.

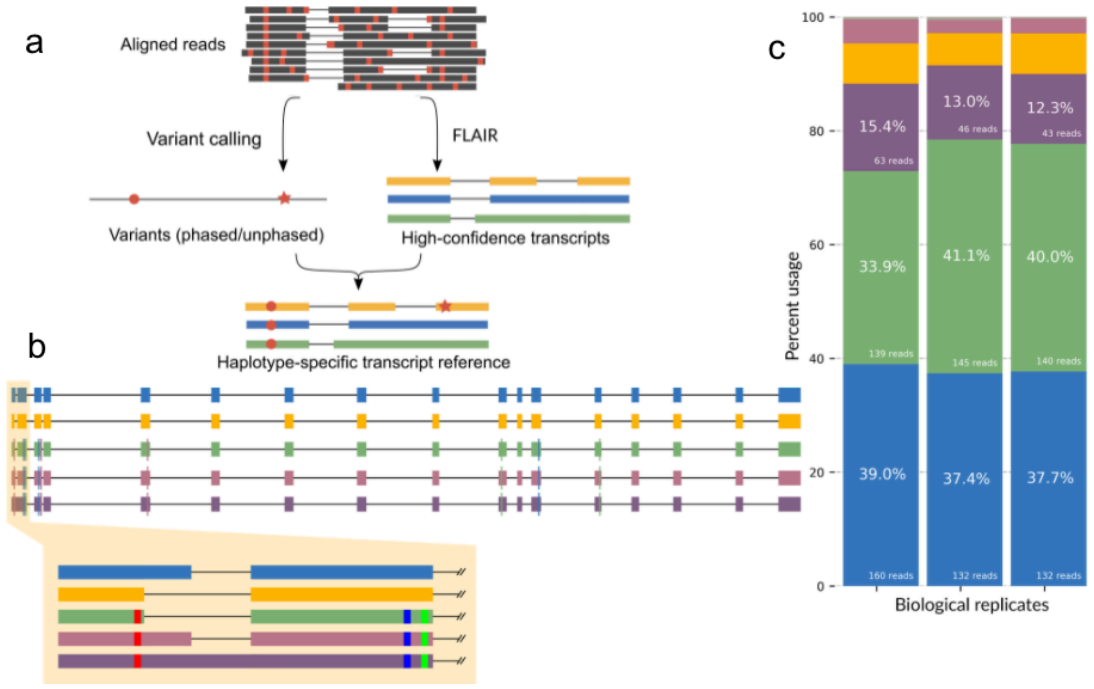

Figure 3.1 Variant-aware transcript detection by FLAIR2. a *Our computational workflow for identifying haplotype-specific transcripts in long reads.* ***b*** *FLAIR transcript models for MCM5 with the highest expression are plotted using different colors for each transcript's exons. The highlighted portion shows alternative splicing and the smaller blocks within exons indicate variants.* ***c*** *Stacked bar chart showing the proportion of transcript expression of transcripts from* ***b*** *as matched by color for each of the replicates sequenced.*

Knockdown of ADAR1 is accompanied by global downregulation of inosines

To improve our understanding of A-to-I editing on the cancer transcriptome, we knocked down ADAR1 followed by short- and long-read RNA sequencing (Fig 3.2a). ADAR1 was knocked down in H1975 cells using siRNAs to achieve 70-80% knockdown of ADAR1 protein levels (Fig 3.2b). We sequenced the three ADAR knockdown and three control knockdown samples with Illumina and nanopore sequencing. We observed 55.1, 73.7, 78.8% decrease in ADAR expression from our normalized Illumina RNA-Seq data, with ADAR being the most downregulated gene (Fig 3.2c). We prepared R2C2 cDNA for each of our samples

and sequenced them in matched ADAR and control knockdown pairs using three MinIONs. We obtained an average of 11.7 gigabases with median read length 9,599 bp from each MinION (Table 3). We report a median accuracy of 99.3% and median read length of 1,287 bp from our consensus-called reads. As the number of consensus-called and demultiplexed reads was less than ideal, we decided to pool all of the samples in each condition together for further analyses.

	Pool 1	Pool 2	Pool 3
Total GB basecalled reads	18.5	7.10	9.56
Number basecalled reads	1,423,603	713,990	778,145
Total GB consensus reads	0.999	0.600	1.03
Median length of consensus reads	1,046	1,192	1,816
Number aligned CTRL KD consensus reads	445,285	267,746	252,193
Number aligned ADAR KD consensus reads	379,472	184,312	169,506
Number aligned CTRL KD size-selected reads	-	-	6,716
Number aligned CTRL KD size-selected reads	-	-	141,754

Table 3 R2C2 nanopore sequencing numbers. *For each ADAR KD and control KD sample pool that was sequenced on a MinION, we report the total number of reads obtained from sequencing after basecalling, consensus calling, and minimap2 alignment to the hg38 genome. We also show the number of gigabases of reads after basecalling and consensus calling, as well as the median length of the consensus reads.*

Inosine detection in short and long reads

We used reditools to catalog nucleotides at every position in the Illumina data and filtered for the positions that conformed to A-to-I expectations (i.e. positions with an A or T in the reference and read support for G or C). We identified 334 A->G mismatches in the Illumina data that were significantly changed upon ADAR knockdown (Methods), with the majority (324)

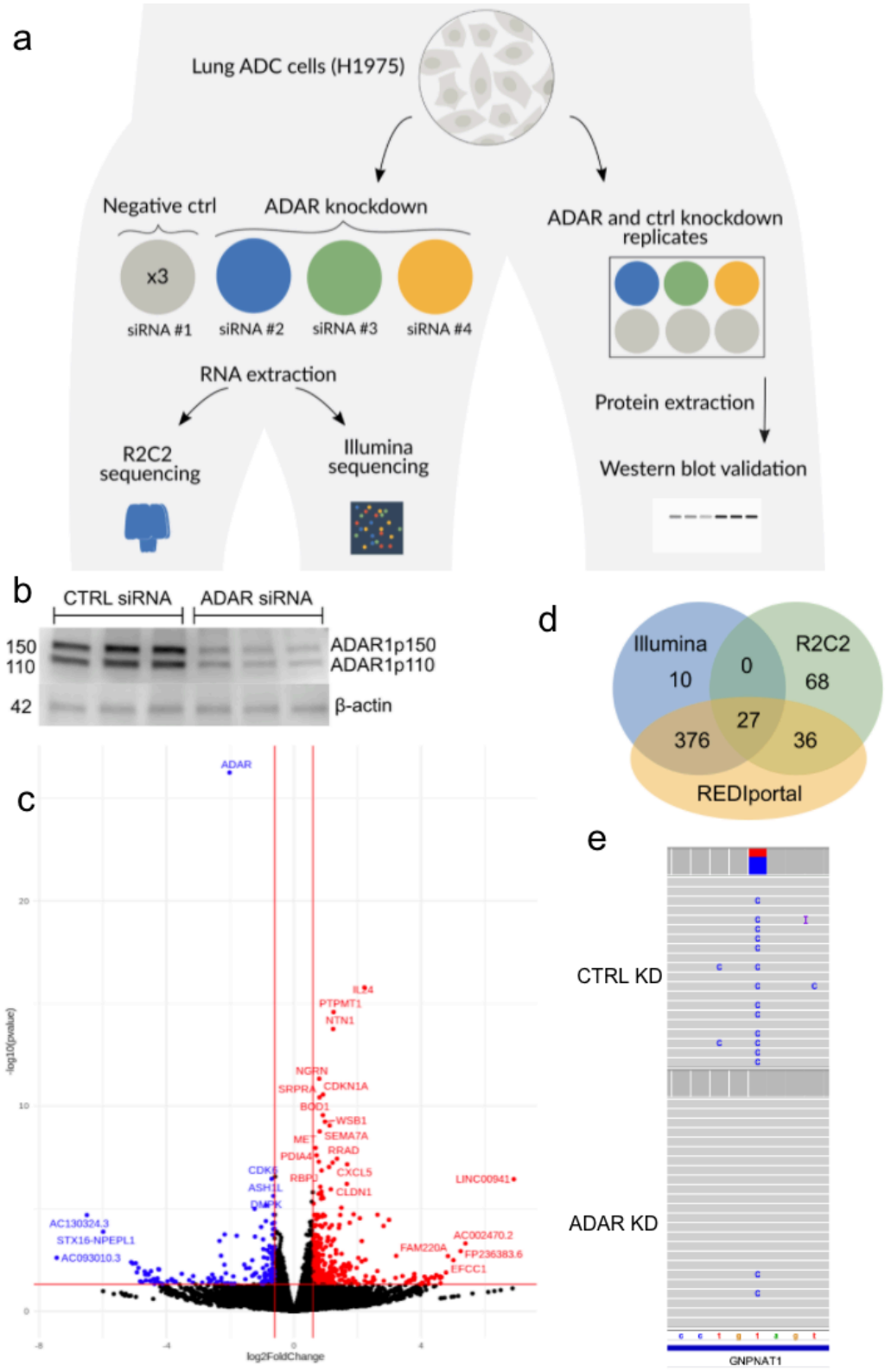
a
Lung ADC cells (H1975)
Negative ctrl
ADAR knockdown
x3
siRNA #1
siRNA #2
siRNA #3
siRNA #4
ADAR and ctrl knockdown replicates
RNA extraction
R2C2 sequencing
Illumina sequencing
Protein extraction
Western blot validation
b
CTRL siRNA
ADAR siRNA
150
110
42
ADAR1p150
ADAR1p110
β-actin
c
ADAR
IL34
PTPMT1
NTN1
NGRN
SRPRA
CDKN1A
BOD1
WSB1
SEMA7A
MET
RRAD
PDIA4
CDK6
RBPJ
CXCL5
ASH1L
CLDN1
DMPK
LINC00941
AC130324.3
STX16-NPEPL1
AC093010.3
AC002470.2
FAM220A
FP236383.6
EFCC1
-log10(pvalue)
log2FoldChange
d
Illumina
R2C2
REDIportal
10
0
68
27
376
36
e
CTRL KD
ADAR KD
GNPNAT1

Figure 3.2 Identification of downregulated inosines with short- and long-read RNA-Seq. a, *Experimental workflow of ADAR knockdown in H1975 cells.* ***b,*** *Western blot.* ***c,*** *Volcano plot of differentially expressed genes identified from Illumina sequencing. Red: genes with increased expressed after ADAR knockdown; blue: genes whose expression went down; black: no change in expression.* ***d,*** *Venn diagram comparison of the significantly downregulated inosines identified with Illumina, R2C2 nanopore, or present in the REDIportal database (hg38 liftover).* ***e,*** *IGV browser view of a downregulated inosine at chr14:52775760 in GNPNAT1 in the R2C2 data.*

of these positions present in the REDIportal database (Mansi et al. 2021). Of these 334 A-to-I events, 312 were downregulated in the knockdown conditions and 12 were upregulated.

We considered longshot and PEPPER-Margin-DeepVariant variant calls to identify an initial set of A-to-G mismatches that we would then reclassify as A-to-I edits with REDIportal and downregulation analyses. Both variant callers identified variants that could be categorized as inosines that the other caller missed; as such, we combined all the variant calls from both tools. Starting with the combined variant calls, we identified 63 significantly changed A-to-I events (Fisher's exact $p<0.1$) with a greater than 10% difference in proportion of edited reads that were also present in REDIportal (Fig 3.2d); as expected, most (62/63) were downregulated in the ADAR knockdown samples. Of the significant nanopore-identified inosines, 27 were also identified as significantly downregulated in the Illumina data (Fig 3.2d,e). Defining type I hyperediting as positions with >40% of adenosine residues edited (Tavakoli et al. 2021) in our control data, we find that approximately half (79/131) of the significantly downregulated inosines were considered type I hyperedited. The inosines identified as significantly differentially edited with nanopore but not in Illumina were generally those that received insufficient coverage of the edited position in the short reads, such as the cases shown in Fig 3.3a and 3.3b. In conclusion, while the quantity of short read data will typically surpass that of longs increasing the number of inosines detected, long reads are advantageous for detecting certain novel A-to-I events.

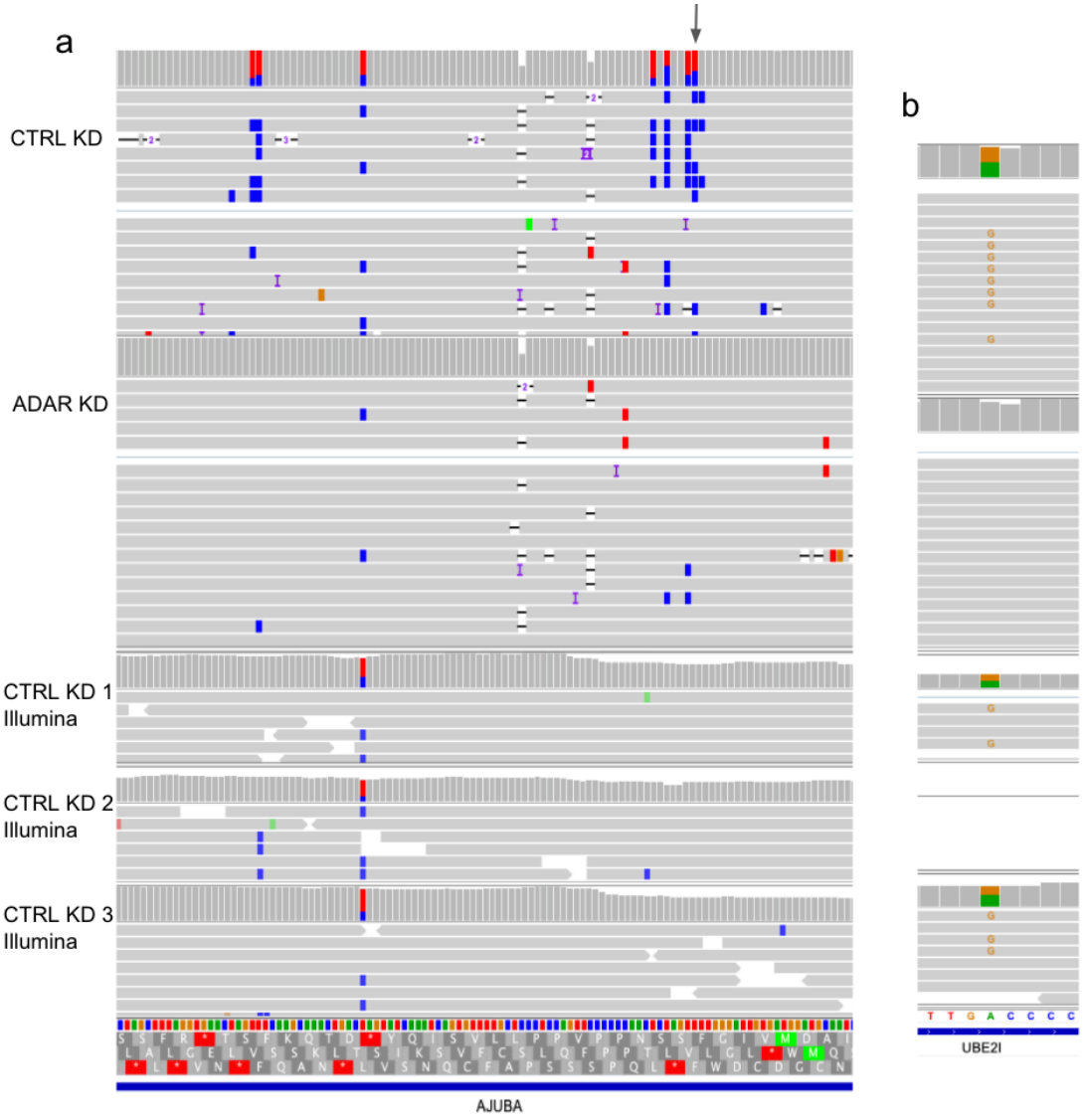

Figure 3.3 Significantly downregulated A-to-I detected with nanopore and not in the Illumina data. *IGV shots of nanopore and Illumina data aligned to hg38.* ***a,*** *Gray arrow indicates the differentially edited position found in nanopore but not Illumina and is a known editing position in REDIportal.* ***b****, Potential A-to-I editing position found with nanopore but not present in REDIportal. There were no reads aligning to UBE2I in the second Illumina CTRL KD replicate.*

Long reads clarify the transcriptional context of inosines

With the Illumina data, we were not able to find many convincing cases of alternative splicing. We ran the differential splicing analyses tools MESA, juncBASE (Brooks et al. 2011), and JUM (Q. Wang and Rio 2018). However, none of the identified splicing events were significant after multiple testing correction. With our nanopore data, we sought to find edits associated with other edits or splicing changes that could be overlooked in the Illumina data due to mapping difficulties or length limitations. We performed a systematic analysis of all inosine-inosine associations within single molecule reads. For each inosine, we looked at the nearest 20 variants, checked all of the reads that overlapped both variants to count the frequency they co-occured with each other, and performed a Fisher's test to discover significantly associated positions. In *MRPL30*, we noted coordinated inosine editing occurring more than 500 bp apart (p=2.35e-6) (Fig 3.4a). The predicted secondary structure of the 3' UTR consists of a hairpin bringing the two sites in closer proximity. We also noticed a pattern in the 3' UTR of melanoregulin (*MREG*) transcripts whereby splicing alterations appeared to be coordinated with A-to-I edits. Our nanopore data show splicing within the 3' UTR of *MREG*, there are several positions proximal to splice sites that are edited and unspliced in the CTRL KD samples (Fig 3.4b). The STAR short-read aligner did not detect these splice junctions in the short reads due to a lack of aligned reads. We then looked for other genes that demonstrated the same mutually exclusive pattern of reads either containing an inosine or having an intron spliced out. We found 145 type I hyperedited sites that resided within introns of other reads assigned to that gene. Three of these sites can be found in the 3'UTR of *CWF19L1* (Fig 3.4c).

Long reads can identify type II hyperediting

ADAR tends to produce clusters of inosines on a transcript, which we define as type II hyperediting (Tavakoli et al. 2021). Type II hyperedited transcripts have been associated with nuclear retention or degradation (Prasanth et al. 2005; Scadden 2005, 2007; Hundley, Krauchuk, and Bass 2008; L.-L. Chen and Carmichael 2009). First, we note a pattern of ADAR

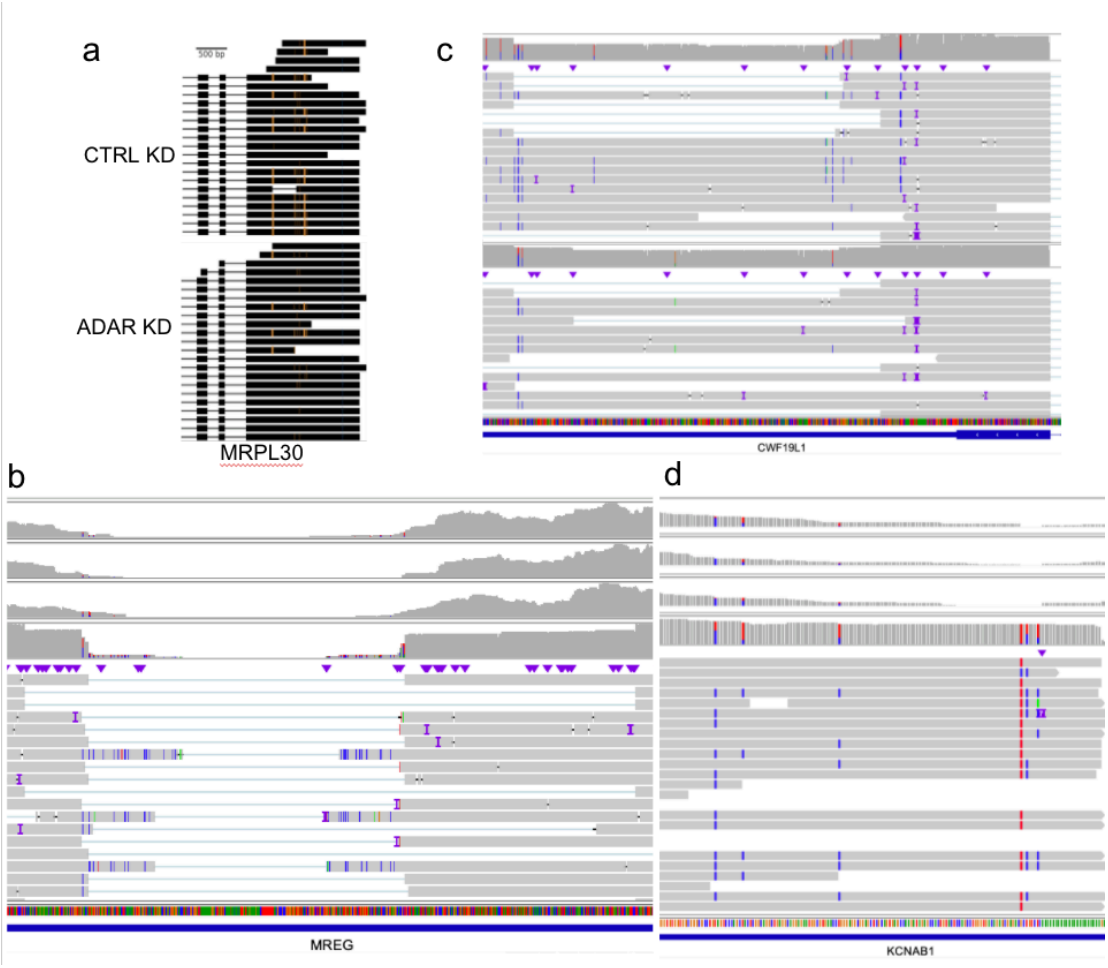

Figure 3.4 Long-range features of inosines observed with nanopore sequencing. *IGV browser views displaying* ***a*** *coordinated editing,* ***b*** *and* ***c*** *disruption of splicing in the presence of editing, and* ***d*** *type II hyperediting. In* ***a*** *and* ***b****, the dataset on top displays the control nanopore reads and the bottom panel displays the ADAR knockdown reads. In* ***b*** *and* ***d,*** *the top three coverage tracks are Illumina CTRL KD samples and the bottom coverage track and reads are displaying the nanopore CTRL KD reads. In* ***a****, orange marks correspond to A->G mismatches and in* ***c, b,*** *and* ***d,*** *positions marked with blue mismatches are T->C mismatches (A->G on the negative strand).*

editing in which transcripts that are edited tend to have multiple edits. The control knockdown data in aggregate show that 38.7% of reads contain at least one edit, and of the reads that are edited, 77.9% contain more than one edit. On detecting multiple edits in short-read RNA-Seq, if the edits are too distant, or if a read contains many mismatches on account of A-to-I

hyperediting (type II), multiply edited reads may not align to the genome and evade detection (Porath, Carmi, and Levanon 2014). With our R2C2 data we were able to identify hyperedited regions with the additional connectivity information. Hyperedited regions were identified as any window that contained at least three A-to-I edits distributed within every 150 bp. To expand our search space, we used the larger set of all inosines found in our nanopore data and REDIportal that were not necessarily significantly downregulated after knockdown as well as the novel significantly downregulated inosines discovered with nanopore only. With this approach, we identified 99 regions that overlapped with known type II hyperediting (Porath, Carmi, and Levanon 2014) as well as 17 novel hyperedited regions with examples in Fig 3.3a and 3,4d.

Discussion

The additive complexity of RNA editing and splicing on the transcriptome, in addition to the disease implications of aberrations in these processes, necessitate methods for more thorough profiling of RNA transcripts. We sought to bridge our understanding of A-to-I editing using short- and long-read sequencing to identify edits more extensively as well as investigate any events that require the full transcriptional context to decipher. We knocked down ADAR in lung adenocarcinoma cells and sequenced the cDNA with the accurate R2C2 nanopore sequencing method. We were able to discover novel type I and type II hyperediting (Fig 3.2e and 3.4d), sites that are coordinated with each other (Fig 3.4a), and sites that may disrupt splicing (Fig 3.4b,c). In this study, we found cases where 3'UTRs were spliced or edited in a mutually exclusive manner. From another study, ADAR-dependent editing of the 3'UTR has been observed to increase expression (Abukar et al. 2021). Elevated levels of editing present in H1975 cells could result in the promoted expression of those transcripts receiving edits in their 3'UTRs.

We were not able to achieve as high a throughput as we expected and thus needed to combine our data from biological replicates. Further studies may benefit from increased

sequencing depth for both short and long reads, since the lack of significant alternatively spliced genes may be due to poor statistical power to detect significance or insufficient levels of knockdown. Also, to capture more intronic A-to-I editing, selection of longer molecules may be necessary to sequence incompletely spliced RNAs on nanopore. Nevertheless, we were still able to build computational pipelines to leverage our accurate nanopore data in ways that surpassed the limitations of short reads, continuing to pave a way for the adoption of long reads for characterizing RNA splicing and editing in cancers.

Methods

Cell culture and siRNA knockdown

H1975 cells were cultured in T75 flasks with DMEM + 10% FBS media. Cells were split 1:4 every 3 days using a 0.25% trypsin 0.52 mM EDTA solution. Trypsin solution was neutralized using an approximately equal volume of media.

For ADAR and control knockdowns, we used Silencer Select siRNAs s1007, s1008, and s1009 for ADAR1 and Silencer Select Negative Control No. 1 at 15 nmol for 72 hours. Cells that would be subject to RNA extraction were cultured in 10 cm dishes. In tandem, cells were plated for western blotting in 6-well plates. Given the vessel, the appropriate amount of siRNA was added to the media when the cells were 80% confluent.

Western blotting

We have uploaded the protocol to protocols.io (Robinson and Tang 2020). Briefly, after siRNA treatment, the media were aspirated off and 200 ul of cold RIPA and proteinase K solution were added to each well. Cells were scraped off, transferred to cold tubes, and centrifuged at 15,000 x g for 7 minutes. Leaving the pellet, the supernatant lysate was then retained and sonicated. Protein lysates were sonicated twice for 30 seconds, with 1 minute on ice in between. Protein concentrations were measured with the Pierce BCA Protein Assay.

Protein lysates were loaded into precast gels. We used ADAR1 primary antibody (ab88574) and goat secondary (ab205719) and imaged on LI-COR C-digit blot scanner.

RNA extraction

Media was aspirated off and the dishes were washed 3x with ice cold dPBS. 1 ml of tri-reagent was added to each dish and cells were scraped off. Cells suspended in tri-reagent were used as input into the Zymo direct-zol kit. Following elution from the direct-zol kit, RNA quality and concentration were evaluated with the nanodrop, qubit, and tapestation.

R2C2 and nanopore sequencing

We followed the R2C2 protocol from Vollmers et. al. (Vollmers et al. 2021). We also have these steps written out on protocols.io. In summary, our steps were as follows: 200 ng of RNA (1 ug for Pools 1 and 2) were reverse transcribed with SmartSeq and barcoded oligo-dTs. RT product underwent lambda exonuclease and RNase A digestion, followed by 15 cycles of PCR using KAPA Hifi HotStart ReadyMix. Next, cDNA was cleaned with 0.8:1 ampure bead purification. Pool 2 was cleaned using Zymo Select-A-Size for fragments larger than 300 nt, adding an extra empty spin step after the second wash. Pool 1 had 4 samples pooled together. Two were size-selected for fragments larger than 3 kb using a low-melt agarose gel extraction. The two samples to be size-selected were first pooled, then run on a 1% low-melt agarose gel made with TAE. A gel slice containing cDNA above 3 kb was cut out and placed in twice the volume of beta-agarase buffer, incubating on ice. The buffer was refreshed after 20 minutes. After another 20 minutes, the buffer was removed and the gel was melted at 65 C for 10 minutes. The gel was then incubated overnight with the addition of 2 ul of beta-agarase per 300 ul of gel. A bead purification was performed on the DNA-containing digested gel. Final R2C2 cDNA concentration was assessed on the nanodrop prior to nanopore sequencing preparation. 200 ng of nanopore library was loaded onto a flow cell at time. Excess library was stored at 4C. After 24 hours, any remaining library was loaded after flushing the flow cell with buffer A (list

components) and DNAse I according to ONT protocol. Reads were basecalled with guppy 4.4.1 and consensus called and demultiplexed using C3POa.

FLAIR splice site fidelity checking

After an ungapped alignment of reads to transcripts, the top transcript alignments for each read as determined by minimap2 mapq score are examined using custom python scripts in FLAIR. We ran FLAIR collapse with both the –stringent and –check_splice parameters to ensure accuracy of read-isoform assignments. The --stringent parameter enforces that 80% of bases match between the read and assigned isoform as well as that the read spans into 25 bp of the first and last exons. The --check_splice parameter enforces that 4 out of 6 bases flanking every splice site in the transcript are matched in a given read and that there are no indels larger than 3 bp at a splice site.

FLAIR novel isoform detection (FLAIR-collapse)

To summarize the unassigned reads into the isoforms they represent with high-confidence, FLAIR first uses minimap2 (H. Li 2016) to align long reads to the genome. The high error rate of nanopore nvRNA or standard cDNA sequencing often results in spurious alignments around splice sites; to combat this, FLAIR then corrects unsupported splice sites with the closest splice site that contains more evidence i.e. splice sites found in annotations or short-read sequencing. The full-length, corrected reads are then grouped by their splice junction chains and FLAIR will call transcription start and end sites for each group, collapsing each group into one or more representative first-pass isoform. Next, FLAIR assigns each read to a first-pass isoform by realigning the reads to the isoforms and identifying the best alignment. The novel FLAIR isoform set arises from filtering the first-pass set for the isoforms that pass a minimum supporting read threshold.

SIRV analysis

We analyzed SIRV reads that aligned with the SIRV1-SIRV7 references from the LRGASP mouse embryonic stem cell R2C2 sequencing replicates. We ran FLAIR2 providing the genome annotation and with the default minimum supporting read count of 3 (-s). We used the -L parameter and supplied a genome annotation for the stringtie2 run. For FLAMES, we used the SIRV config file with a minimum supporting read count of 3. We used gffcompare (Pertea and Pertea 2020) to calculate transcript-level sensitivity and precision of each tool's transcript reference with the ground truth, using a wiggle room of 50 bp at the transcription start sites and terminal ends for matching (-e 50 and -d 50).

FLAIR-variant

Criteria for high sequence identity with an isoform are based on the stringent criteria from Workman et al. For multi-exonic isoform assignments, high sequence identity near the bases that flank the splice sites is another requirement. If the read alignment contains deletions near any of the splice sites or insertions Ties between assignments were broken using alignments with fewer softclipped bases at the ends of the reads and minimizing the number of unmapped bases on the transcript.

Illumina A-to-I analysis

REDItools was used to tabulate the number of reads supporting each base at every position. The REDItools output was filtered using custom python scripts for positions that contained guanosine mismatches at positions where the reference base was an adenosine for genes corresponding to the forward strand of the genome, and the reverse complement for those on the reverse strand. Positions with less than 15% putative editing were filtered out. The counts of the reference and alternate allele in each of the samples for the remaining positions were supplied to DRIMSeq (Nowicka and Robinson 2016) for differential testing between two conditions, with the settings that at least 5 reads contained editing (G mismatch) in a minimum of two samples, as well as a coverage of 15 reads minimum in at least 3 samples.

Inosine detection in long reads

We used the pysam python package's pileup method to count A->G or T->C reads at variant positions. Next, we combined our nanopore data by knockdown condition, followed by filtering for positions that had a minimum coverage of 10 in either condition and a change in percentage of edited reads after ADAR knockdown of 10% or more. We performed a Fisher's exact test to assess the significance of the A-to-I differences.

Inosine coordination analysis for long reads

We filtered for sites that were type I hyperedited (i.e. more than 40% of residues were edited) and had at least 10 reads that were edited. We also required that at least 10 reads had the edited position spliced out.

Discussion

Through the pairing of novel sequencing methods and development of compatible computational methods, we have demonstrated a way to improve transcriptome analyses for RNA splicing and editing. We have used nanopore cDNA and direct RNA sequencing for the detection of splice variants with full-length molecules, further informing us on the nature of an altered retained intron landscape in chronic lymphocytic leukemia and establishing a relationship between retained introns and poly(A)+ tail length. With nanopore R2C2 sequencing, we were able to leverage the increased single molecule accuracy to detect the long-range effects of A-to-I editing in lung adenocarcinoma tumorigenesis. Throughout these research ventures, we incrementally improved FLAIR's algorithms for long-read RNA analyses, contributing to a tool space for other groups' research use as well.

Finding these RNA processing patterns with long reads is an essential step toward a more complete picture of these cancer transcriptomes, expanding upon the decades of research that preceded long reads. As the technology continues to advance, the additional information afforded by long reads for disentangling longer-range interactions, such as regions with coordinated splicing that short reads cannot span or repetitive regions, becomes increasingly clear. Based on the work that we have done, future studies looking to examine RNA splicing, editing, or modifications could consider a long-read approach. Furthermore, the field can benefit from long reads applied to single-cell or spatial transcriptomics to elucidate the heterogeneity of cancers and lend new perspectives to cancer drug resistance and the tumor microenvironment. While there are still many areas to be explored in depth with long-read RNA-Seq, this higher resolution approach will inevitably bring forth the greater understanding we need to progress the behemoth that is cancer prevention, diagnosis, treatment, and management.

References

Abukar, Asra, Martin Wipplinger, Ananya Hariharan, Suna Sun, Manuel Ronner, Marika Sculco, Agata Okonska, et al. 2021. “Double-Stranded RNA Structural Elements Holding the Key to Translational Regulation in Cancer: The Case of Editing in RNA-Binding Motif Protein 8A.” *Cells* 10 (12). https://doi.org/10.3390/cells10123543.

Adams, M. D., J. M. Kelley, J. D. Gocayne, M. Dubnick, M. H. Polymeropoulos, H. Xiao, C. R. Merril, A. Wu, B. Olde, and R. F. Moreno. 1991. “Complementary DNA Sequencing: Expressed Sequence Tags and Human Genome Project.” *Science* 252 (5013): 1651–56.

Alsafadi, Samar, Alexandre Houy, Aude Battistella, Tatiana Popova, Michel Wassef, Emilie Henry, Franck Tirode, et al. 2016. “Cancer-Associated SF3B1 Mutations Affect Alternative Splicing by Promoting Alternative Branchpoint Usage.” *Nature Communications* 7 (February): 10615.

Amin, Elianna M., Yuan Liu, Su Deng, Kay See Tan, Neel Chudgar, Marty W. Mayo, Francisco Sanchez-Vega, Prasad S. Adusumilli, Nikolaus Schultz, and David R. Jones. 2017. “The RNA-Editing Enzyme ADAR Promotes Lung Adenocarcinoma Migration and Invasion by Stabilizing FAK.” *Science Signaling* 10 (497). https://doi.org/10.1126/scisignal.aah3941.

Anvar, Seyed Yahya, Guy Allard, Elizabeth Tseng, Gloria M. Sheynkman, Eleonora de Klerk, Martijn Vermaat, Raymund H. Yin, et al. 2018. “Full-Length mRNA Sequencing Uncovers a Widespread Coupling between Transcription Initiation and mRNA Processing.” *Genome Biology* 19 (1): 46.

Athanasiadis, Alekos, Alexander Rich, and Stefan Maas. 2004. “Widespread A-to-I RNA Editing of Alu-Containing mRNAs in the Human Transcriptome.” *PLoS Biology* 2 (12): e391.

Bajad, Prajakta, Michael F. Jantsch, Liam Keegan, and Mary O’Connell. 2017. “A to I Editing in Disease Is Not Fake News.” *RNA Biology* 14 (9): 1223–31.

Baltimore, David. 1970. “Viral RNA-Dependent DNA Polymerase: RNA-Dependent DNA Polymerase in Virions of RNA Tumour Viruses.” *Nature* 226 (June): 1209.

Baralle, Francisco E., and Jimena Giudice. 2017. “Alternative Splicing as a Regulator of Development and Tissue Identity.” *Nature Reviews. Molecular Cell Biology* 18 (7): 437–51.

Bazak, Lily, Ami Haviv, Michal Barak, Jasmine Jacob-Hirsch, Patricia Deng, Rui Zhang, Farren J. Isaacs, et al. 2014. “A-to-I RNA Editing Occurs at over a Hundred Million Genomic Sites, Located in a Majority of Human Genes.” *Genome Research* 24 (3): 365–76.

Bernstein, Bradley E., Michael Kamal, Kerstin Lindblad-Toh, Stefan Bekiranov, Dione K. Bailey, Dana J. Huebert, Scott McMahon, et al. 2005. “Genomic Maps and Comparative Analysis of Histone Modifications in Human and Mouse.” *Cell* 120 (2): 169–81.

Bolisetty, Mohan T., Gopinath Rajadinakaran, and Brenton R. Graveley. 2015. “Determining Exon Connectivity in Complex mRNAs by Nanopore Sequencing.” *Genome Biology* 16 (September): 204.

Bradley, Robert K., Jason Merkin, Nicole J. Lambert, and Christopher B. Burge. 2012. “Alternative Splicing of RNA Triplets Is Often Regulated and Accelerates Proteome Evolution.” *PLoS Biology* 10 (1): e1001229.

Braun, Christian J., Monica Stanciu, Paul L. Boutz, Jesse C. Patterson, David Calligaris, Fumi Higuchi, Rachit Neupane, et al. 2017. “Coordinated Splicing of Regulatory Detained Introns within Oncogenic Transcripts Creates an Exploitable Vulnerability in Malignant Glioma.” *Cancer Cell* 32 (4): 411–26.e11.

Bresson, Stefan M., Olga V. Hunter, Allyson C. Hunter, and Nicholas K. Conrad. 2015. “Canonical Poly(A) Polymerase Activity Promotes the Decay of a Wide Variety of Mammalian Nuclear RNAs.” *PLoS Genetics* 11 (10): e1005610.

Brooks, Angela N., Li Yang, Michael O. Duff, Kasper D. Hansen, Jung W. Park, Sandrine Dudoit, Steven E. Brenner, and Brenton R. Graveley. 2011. “Conservation of an RNA Regulatory Map between Drosophila and Mammals.” *Genome Research* 21 (2): 193–202.

Brown, Carolyn J., Andrea Ballabio, James L. Rupert, Ronald G. Lafreniere, Markus Grompe, Rossana Tonlorenzi, and Huntington F. Willard. 1991. “A Gene from the Region of the

Human X Inactivation Centre Is Expressed Exclusively from the Inactive X Chromosome." *Nature* 349 (January): 38.

Bueno, Raphael, Eric W. Stawiski, Leonard D. Goldstein, Steffen Durinck, Assunta De Rienzo, Zora Modrusan, Florian Gnad, et al. 2016. "Comprehensive Genomic Analysis of Malignant Pleural Mesothelioma Identifies Recurrent Mutations, Gene Fusions and Splicing Alterations." *Nature Genetics* 48 (4): 407–16.

Burnashev, N., H. Monyer, P. H. Seeburg, and B. Sakmann. 1992. "Divalent Ion Permeability of AMPA Receptor Channels Is Dominated by the Edited Form of a Single Subunit." *Neuron* 8 (1): 189–98.

Byrne, Ashley, Anna E. Beaudin, Hugh E. Olsen, Miten Jain, Charles Cole, Theron Palmer, Rebecca M. DuBois, E. Camilla Forsberg, Mark Akeson, and Christopher Vollmers. 2017. "Nanopore Long-Read RNAseq Reveals Widespread Transcriptional Variation among the Surface Receptors of Individual B Cells." *Nature Communications* 8 (July): 16027.

Byrne, Ashley, Megan A. Supple, Roger Volden, Kristin L. Laidre, Beth Shapiro, and Christopher Vollmers. 2019. "Depletion of Hemoglobin Transcripts and Long Read Sequencing Improves the Transcriptome Annotation of the Polar Bear (Ursus Maritimus)." *bioRxiv*. https://doi.org/10.1101/527978.

Carrocci, Tucker J., Douglas M. Zoerner, Joshua C. Paulson, and Aaron A. Hoskins. 2017. "SF3b1 Mutations Associated with Myelodysplastic Syndromes Alter the Fidelity of Branchsite Selection in Yeast." *Nucleic Acids Research* 45 (8): 4837–52.

Chang, Hyeshik, Jaechul Lim, Minju Ha, and V. Narry Kim. 2014. "TAIL-Seq: Genome-Wide Determination of poly(A) Tail Length and 3' End Modifications." *Molecular Cell* 53 (6): 1044–52.

Chen, C. X., D. S. Cho, Q. Wang, F. Lai, K. C. Carter, and K. Nishikura. 2000. "A Third Member of the RNA-Specific Adenosine Deaminase Gene Family, ADAR3, Contains Both Single- and Double-Stranded RNA Binding Domains." *RNA* 6 (5): 755–67.

Chen, Leilei, Yan Li, Chi Ho Lin, Tim Hon Man Chan, Raymond Kwok Kei Chow, Yangyang Song, Ming Liu, et al. 2013. “Recoding RNA Editing of AZIN1 Predisposes to Hepatocellular Carcinoma.” *Nature Medicine* 19 (2): 209–16.

Chen, Ling-Ling, and Gordon G. Carmichael. 2009. “Altered Nuclear Retention of mRNAs Containing Inverted Repeats in Human Embryonic Stem Cells: Functional Role of a Nuclear Noncoding RNA.” *Molecular Cell* 35 (4): 467–78.

Cho, Hyunghoon, Joe Davis, Xin Li, Kevin S. Smith, Alexis Battle, and Stephen B. Montgomery. 2014. “High-Resolution Transcriptome Analysis with Long-Read RNA Sequencing.” *PloS One* 9 (9): e108095.

Cloonan, Nicole, Alistair R. R. Forrest, Gabriel Kolle, Brooke B. A. Gardiner, Geoffrey J. Faulkner, Mellissa K. Brown, Darrin F. Taylor, et al. 2008. “Stem Cell Transcriptome Profiling via Massive-Scale mRNA Sequencing.” *Nature Methods* 5 (7): 613–19.

Corvelo, André, Martina Hallegger, Christopher W. J. Smith, and Eduardo Eyras. 2010. “Genome-Wide Association between Branch Point Properties and Alternative Splicing.” *PLoS Computational Biology* 6 (11): e1001016.

Darman, Rachel B., Michael Seiler, Anant A. Agrawal, Kian H. Lim, Shouyong Peng, Daniel Aird, Suzanna L. Bailey, et al. 2015. “Cancer-Associated SF3B1 Hotspot Mutations Induce Cryptic 3’ Splice Site Selection through Use of a Different Branch Point.” *Cell Reports* 13 (5): 1033–45.

Deamer, David, Mark Akeson, and Daniel Branton. 2016. “Three Decades of Nanopore Sequencing.” *Nature Biotechnology* 34 (5): 518–24.

DeBoever, Christopher, Emanuela M. Ghia, Peter J. Shepard, Laura Rassenti, Christian L. Barrett, Kristen Jepsen, Catriona H. M. Jamieson, Dennis Carson, Thomas J. Kipps, and Kelly A. Frazer. 2015. “Transcriptome Sequencing Reveals Potential Mechanism of Cryptic 3’ Splice Site Selection in SF3B1-Mutated Cancers.” *PLoS Computational Biology* 11 (3): e1004105.

Deveson, Ira W., Marion E. Brunck, James Blackburn, Elizabeth Tseng, Ting Hon, Tyson A. Clark, Michael B. Clark, et al. 2018. “Universal Alternative Splicing of Noncoding Exons.” *Cell Systems* 6 (2): 245–55.e5.

Dong, Chuanpeng, Annamaria Cesarano, Giuseppe Bombaci, Jill L. Reiter, Christina Y. Yu, Yue Wang, Zhaoyang Jiang, et al. 2021. “Intron Retention-Induced Neoantigen Load Correlates with Unfavorable Prognosis in Multiple Myeloma.” *Oncogene* 40 (42): 6130–38.

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

Eberle, Michael A., Epameinondas Fritzilas, Peter Krusche, Morten Källberg, Benjamin L. Moore, Mitchell A. Bekritsky, Zamin Iqbal, et al. 2016. “A Reference Data Set of 5.4 Million Phased Human Variants Validated by Genetic Inheritance from Sequencing a Three-Generation 17-Member Pedigree.” *Genome Research*, November. https://doi.org/10.1101/gr.210500.116.

Eckmann, Christian R., Christiane Rammelt, and Elmar Wahle. 2011. “Control of poly(A) Tail Length.” *Wiley Interdisciplinary Reviews. RNA* 2 (3): 348–61.

Edge, Peter, Vineet Bafna, and Vikas Bansal. 2017. “HapCUT2: Robust and Accurate Haplotype Assembly for Diverse Sequencing Technologies.” *Genome Research* 27 (5): 801–12.

Edge, Peter, and Vikas Bansal. 2019. “Longshot Enables Accurate Variant Calling in Diploid Genomes from Single-Molecule Long Read Sequencing.” *Nature Communications* 10 (1): 4660.

Ernst, Jason, and Manolis Kellis. 2010. “Discovery and Characterization of Chromatin States for Systematic Annotation of the Human Genome.” *Nature Biotechnology* 28 (8): 817–25.

Ernst, Jason, Pouya Kheradpour, Tarjei S. Mikkelsen, Noam Shoresh, Lucas D. Ward, Charles B. Epstein, Xiaolan Zhang, et al. 2011. “Mapping and Analysis of Chromatin State Dynamics in Nine Human Cell Types.” *Nature* 473 (7345): 43–49.

Filichkin, Sergei A., and Todd C. Mockler. 2012. “Unproductive Alternative Splicing and Nonsense mRNAs: A Widespread Phenomenon among Plant Circadian Clock Genes.” *Biology Direct* 7 (July): 20.

Furney, Simon J., Malin Pedersen, David Gentien, Amaury G. Dumont, Audrey Rapinat, Laurence Desjardins, Samra Turajlic, et al. 2013. “SF3B1 Mutations Are Associated with Alternative Splicing in Uveal Melanoma.” *Cancer Discovery* 3 (10): 1122–29.

Fu, Xiang-Dong. 2017. “Exploiting the Hidden Treasure of Detained Introns.” *Cancer Cell* 32 (4): 393–95.

Fu, Xing, Ming Tian, Jia Gu, Teng Cheng, Ding Ma, Ling Feng, and Xing Xin. 2017. “SF3B1 Mutation Is a Poor Prognostic Indicator in Luminal B and Progesterone Receptor-Negative Breast Cancer Patients.” *Oncotarget* 8 (70): 115018–27.

Garalde, Daniel R., Elizabeth A. Snell, Daniel Jachimowicz, Botond Sipos, Joseph H. Lloyd, Mark Bruce, Nadia Pantic, et al. 2018. “Highly Parallel Direct RNA Sequencing on an Array of Nanopores.” *Nature Methods*, January. https://doi.org/10.1038/nmeth.4577.

Garraway, Levi A., and Eric S. Lander. 2013. “Lessons from the Cancer Genome.” *Cell* 153 (1): 17–37.

Glinos, Dafni A., Garrett Garborcauskas, Paul Hoffman, Nava Ehsan, Lihua Jiang, Alper Gokden, Xiaoguang Dai, et al. 2021. “Transcriptome Variation in Human Tissues Revealed by Long-Read Sequencing.” *bioRxiv*. https://doi.org/10.1101/2021.01.22.427687.

Gonzàlez-Porta, Mar, Adam Frankish, Johan Rung, Jennifer Harrow, and Alvis Brazma. 2013. “Transcriptome Analysis of Human Tissues and Cell Lines Reveals One Dominant Transcript per Gene.” *Genome Biology* 14 (7): R70.

Gozani, O., R. Feld, and R. Reed. 1996. “Evidence That Sequence-Independent Binding of Highly Conserved U2 snRNP Proteins Upstream of the Branch Site Is Required for Assembly of Spliceosomal Complex A.” *Genes & Development* 10 (2): 233–43.

Gozani, O., J. Potashkin, and R. Reed. 1998. “A Potential Role for U2AF-SAP 155 Interactions in Recruiting U2 snRNP to the Branch Site.” *Molecular and Cellular Biology* 18 (8): 4752–60.

Gumireddy, Kiranmai, Anping Li, Andrew V. Kossenkov, Masayuki Sakurai, Jinchun Yan, Yan Li, Hua Xu, et al. 2016. “The mRNA-Edited Form of GABRA3 Suppresses GABRA3-Mediated Akt Activation and Breast Cancer Metastasis.” *Nature Communications* 7 (February): 10715.

Han, Leng, Lixia Diao, Shuangxing Yu, Xiaoyan Xu, Jie Li, Rui Zhang, Yang Yang, et al. 2015. “The Genomic Landscape and Clinical Relevance of A-to-I RNA Editing in Human Cancers.” *Cancer Cell* 28 (4): 515–28.

Harbour, J. William, Elisha D. O. Roberson, Hima Anbunathan, Michael D. Onken, Lori A. Worley, and Anne M. Bowcock. 2013. “Recurrent Mutations at Codon 625 of the Splicing Factor SF3B1 in Uveal Melanoma.” *Nature Genetics* 45 (2): 133–35.

Hinrichs, A. S., D. Karolchik, R. Baertsch, G. P. Barber, G. Bejerano, H. Clawson, M. Diekhans, et al. 2006. “The UCSC Genome Browser Database: Update 2006.” *Nucleic Acids Research* 34 (Database issue): D590–98.

Hsiao, Yun-Hua Esther, Jae Hoon Bahn, Yun Yang, Xianzhi Lin, Stephen Tran, Ei-Wen Yang, Giovanni Quinones-Valdez, and Xinshu Xiao. 2018. “RNA Editing in Nascent RNA Affects Pre-mRNA Splicing.” *Genome Research* 28 (6): 812–23.

Hundley, Heather A., Ammie A. Krauchuk, and Brenda L. Bass. 2008. “C. Elegans and H. Sapiens mRNAs with Edited 3’ UTRs Are Present on Polysomes.” *RNA* 14 (10): 2050–60.

Inoue, Daichi, Guo-Liang Chew, Bo Liu, Brittany C. Michel, Joseph Pangallo, Andrew R. D’Avino, Tyler Hitchman, et al. 2019. “Spliceosomal Disruption of the Non-Canonical BAF Complex in Cancer.” *Nature*, October. https://doi.org/10.1038/s41586-019-1646-9.

Jacob, Aishwarya G., and Christopher W. J. Smith. 2017. “Intron Retention as a Component of Regulated Gene Expression Programs.” *Human Genetics* 136 (9): 1043–57.

Jain, Miten, Sergey Koren, Karen H. Miga, Josh Quick, Arthur C. Rand, Thomas A. Sasani, John R. Tyson, et al. 2018. “Nanopore Sequencing and Assembly of a Human Genome with Ultra-Long Reads.” *Nature Biotechnology*, January. https://doi.org/10.1038/nbt.4060.

Jain, Miten, Hugh E. Olsen, Benedict Paten, and Mark Akeson. 2016. “The Oxford Nanopore MinION: Delivery of Nanopore Sequencing to the Genomics Community.” *Genome Biology* 17 (1): 239.

Jain, Miten, Hugh E. Olsen, Daniel J. Turner, David Stoddart, Kira V. Bulazel, Benedict Paten, David Haussler, Huntington F. Willard, Mark Akeson, and Karen H. Miga. 2018. “Linear Assembly of a Human Centromere on the Y Chromosome.” *Nature Biotechnology*, March. https://doi.org/10.1038/nbt.4109.

Jain, Miten, John R. Tyson, Matthew Loose, Camilla L. C. Ip, David A. Eccles, Justin O’Grady, Sunir Malla, et al. 2017. “MinION Analysis and Reference Consortium: Phase 2 Data Release and Analysis of R9.0 Chemistry.” *F1000Research* 6 (May): 760.

Jenjaroenpun, Piroon, Thidathip Wongsurawat, Rui Pereira, Preecha Patumcharoenpol, David W. Ussery, Jens Nielsen, and Intawat Nookaew. 2018. “Complete Genomic and Transcriptional Landscape Analysis Using Third-Generation Sequencing: A Case Study of Saccharomyces Cerevisiae CEN.PK113-7D.” *Nucleic Acids Research*, January. https://doi.org/10.1093/nar/gky014.

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.

Kandoth, Cyriac, Michael D. McLellan, Fabio Vandin, Kai Ye, Beifang Niu, Charles Lu, Mingchao Xie, et al. 2013. “Mutational Landscape and Significance across 12 Major Cancer Types.” *Nature* 502 (7471): 333–39.

Kawahara, Yukio, Shin Kwak, Hui Sun, Kyoko Ito, Hideji Hashida, Hitoshi Aizawa, Seon-Yong Jeong, and Ichiro Kanazawa. 2003. “Human Spinal Motoneurons Express Low Relative Abundance of GluR2 mRNA: An Implication for Excitotoxicity in ALS: AMPA Subunit Expression Profile in Human CNS.” *Journal of Neurochemistry* 85 (3): 680–89.

Kent, W. James. 2002. “BLAT--the BLAST-like Alignment Tool.” *Genome Research* 12 (4): 656–64.

Kesarwani, A. K., O. Ramirez, A. K. Gupta, X. Yang, T. Murthy, A. C. Minella, and M. M. Pillai. 2017. “Cancer-Associated SF3B1 Mutants Recognize Otherwise Inaccessible Cryptic 3’ Splice Sites within RNA Secondary Structures.” *Oncogene* 36 (8): 1123–33.

Kingan, Sarah B., Julie Urban, Christine C. Lambert, Primo Baybayan, Anna K. Childers, Brad S. Coates, Brian Scheffler, Kevin Hackett, Jonas Korlach, and Scott M. Geib. 2019. “A High-Quality Genome Assembly from a Single, Field-Collected Spotted Lanternfly (Lycorma Delicatula) Using the PacBio Sequel II System.” *bioRxiv*. https://doi.org/10.1101/627679.

Kiran, Anmol M., John J. O’Mahony, Komal Sanjeev, and Pavel V. Baranov. 2013. “Darned in 2013: Inclusion of Model Organisms and Linking with Wikipedia.” *Nucleic Acids Research* 41 (Database issue): D258–61.

Križanovic, Krešimir, Amina Echchiki, Julien Roux, and Mile Šikic. 2018. “Evaluation of Tools for Long Read RNA-Seq Splice-Aware Alignment.” *Bioinformatics* 34 (5): 748–54.

Landau, Dan A., Eugen Tausch, Amaro N. Taylor-Weiner, Chip Stewart, Johannes G. Reiter, Jasmin Bahlo, Sandra Kluth, et al. 2015. “Mutations Driving CLL and Their Evolution in Progression and Relapse.” *Nature* 526 (7574): 525–30.

Lazzari, Elisa, Phoebe K. Mondala, Nathaniel Delos Santos, Amber C. Miller, Gabriel Pineda, Qingfei Jiang, Heather Leu, et al. 2017. “Alu-Dependent RNA Editing of GLI1 Promotes Malignant Regeneration in Multiple Myeloma.” *Nature Communications* 8 (1): 1922.

Lee, Mihye, Boseon Kim, and V. Narry Kim. 2014. “Emerging Roles of RNA Modification: m(6)A and U-Tail.” *Cell* 158 (5): 980–87.

Levanon, Erez Y., Eli Eisenberg, Rodrigo Yelin, Sergey Nemzer, Martina Hallegger, Ronen Shemesh, Zipora Y. Fligelman, et al. 2004. “Systematic Identification of Abundant A-to-I Editing Sites in the Human Transcriptome.” *Nature Biotechnology* 22 (8): 1001–5.

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.

Licht, Konstantin, Utkarsh Kapoor, Elisa Mayrhofer, and Michael F. Jantsch. 2016. “Adenosine to Inosine Editing Frequency Controlled by Splicing Efficiency.” *Nucleic Acids Research* 44 (13): 6398–6408.

Li, Heng. 2016. “Minimap and Miniasm: Fast Mapping and de Novo Assembly for Noisy Long Sequences.” *Bioinformatics* 32 (14): 2103–10.

———. 2018. “Minimap2: Pairwise Alignment for Nucleotide Sequences.” *Bioinformatics* , May. https://doi.org/10.1093/bioinformatics/bty191.

Liu, Bo, Yadong Liu, Junyi Li, Hongzhe Guo, Tianyi Zang, and Yadong Wang. 2019. “deSALT: Fast and Accurate Long Transcriptomic Read Alignment with de Bruijn Graph-Based Index.” *Genome Biology* 20 (1): 274.

Liu, Nian, and Tao Pan. 2016. “N6-Methyladenosine–encoded Epitranscriptomics.” *Nature Structural & Molecular Biology* 23 (2): 98–102.

Livingston, John H., Jean-Pierre Lin, Russell C. Dale, Deepak Gill, Paul Brogan, Arnold Munnich, Manju A. Kurian, et al. 2014. “A Type I Interferon Signature Identifies Bilateral Striatal Necrosis due to Mutations in ADAR1.” *Journal of Medical Genetics* 51 (2): 76–82.

Li, Xiaoyu, Xushen Xiong, and Chengqi Yi. 2016. “Epitranscriptome Sequencing Technologies: Decoding RNA Modifications.” *Nature Methods* 14 (1): 23–31.

Lupu, Cristina, Hua Zhu, Narcis I. Popescu, Jonathan D. Wren, and Florea Lupu. 2011. “Novel Protein ADTRP Regulates TFPI Expression and Function in Human Endothelial Cells in Normal Conditions and in Response to Androgen.” *Blood* 118 (16): 4463–71.

Maguire, Sarah L., Andri Leonidou, Patty Wai, Caterina Marchiò, Charlotte Ky Ng, Anna Sapino, Anne-Vincent Salomon, Jorge S. Reis-Filho, Britta Weigelt, and Rachael C. Natrajan. 2015. “SF3B1 Mutations Constitute a Novel Therapeutic Target in Breast Cancer.” *The Journal of Pathology* 235 (4): 571–80.

Malcovati, Luca, Elli Papaemmanuil, David T. Bowen, Jacqueline Boultwood, Matteo G. Della Porta, Cristiana Pascutto, Erica Travaglino, et al. 2011. “Clinical Significance of SF3B1 Mutations in Myelodysplastic Syndromes and Myelodysplastic/myeloproliferative Neoplasms.” *Blood* 118 (24): 6239–46.

Mansi, Luigi, Marco Antonio Tangaro, Claudio Lo Giudice, Tiziano Flati, Eli Kopel, Amos Avraham Schaffer, Tiziana Castrignanò, Giovanni Chillemi, Graziano Pesole, and Ernesto Picardi. 2021. “REDIportal: Millions of Novel A-to-I RNA Editing Events from Thousands of RNAseq Experiments.” *Nucleic Acids Research* 49 (D1): D1012–19.

Martin, Marcel, Lars Maßhöfer, Petra Temming, Sven Rahmann, Claudia Metz, Norbert Bornfeld, Johannes van de Nes, et al. 2013. “Exome Sequencing Identifies Recurrent Somatic Mutations in EIF1AX and SF3B1 in Uveal Melanoma with Disomy 3.” *Nature Genetics* 45 (8): 933–36.

Miga, Karen H., Sergey Koren, Arang Rhie, Mitchell R. Vollger, Ariel Gershman, Andrey Bzikadze, Shelise Brooks, et al. 2020. “Telomere-to-Telomere Assembly of a Complete Human X Chromosome.” *Nature* 585 (7823): 79–84.

Nakamura, Kensuke, Taku Oshima, Takuya Morimoto, Shun Ikeda, Hirofumi Yoshikawa, Yuh Shiwa, Shu Ishikawa, et al. 2011. “Sequence-Specific Error Profile of Illumina Sequencers.” *Nucleic Acids Research* 39 (13): e90.

Nellore, Abhinav, Andrew E. Jaffe, Jean-Philippe Fortin, José Alquicira-Hernández, Leonardo Collado-Torres, Siruo Wang, Robert A. Phillips III, et al. 2016. “Human Splicing Diversity and the Extent of Unannotated Splice Junctions across Human RNA-Seq Samples on the Sequence Read Archive.” *Genome Biology* 17 (1): 266.

Nishikura, Kazuko. 2010. “Functions and Regulation of RNA Editing by ADAR Deaminases.” *Annual Review of Biochemistry* 79: 321–49.

Nowicka, Malgorzata, and Mark D. Robinson. 2016. “DRIMSeq: A Dirichlet-Multinomial Framework for Multivariate Count Outcomes in Genomics.” *F1000Research* 5 (June): 1356.

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

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.

Papaemmanuil, E., M. Cazzola, J. Boultwood, L. Malcovati, P. Vyas, D. Bowen, A. Pellagatti, et al. 2011. “Somatic SF3B1 Mutation in Myelodysplasia with Ring Sideroblasts.” *The New England Journal of Medicine* 365 (15): 1384–95.

Parker, Roy, and Haiwei Song. 2004. “The Enzymes and Control of Eukaryotic mRNA Turnover.” *Nature Structural & Molecular Biology* 11 (2): 121–27.

Payne, Alex, Nadine Holmes, Vardhman Rakyan, and Matthew Loose. 2018. “Whale Watching with BulkVis: A Graphical Viewer for Oxford Nanopore Bulk fast5 Files.” *bioRxiv*. https://doi.org/10.1101/312256.

Pereira, Bernard, Suet-Feung Chin, Oscar M. Rueda, Hans-Kristian Moen Vollan, Elena Provenzano, Helen A. Bardwell, Michelle Pugh, et al. 2016. “The Somatic Mutation Profiles of

2,433 Breast Cancers Refine Their Genomic and Transcriptomic Landscapes." *Nature Communications* 7 (May): 11479.

Pertea, Geo, and Mihaela Pertea. 2020. "GFF Utilities: GffRead and GffCompare." *F1000Research* 9 (304): 304.

Picardi, Ernesto, Caterina Manzari, Francesca Mastropasqua, Italia Aiello, Anna Maria D'Erchia, and Graziano Pesole. 2015. "Profiling RNA Editing in Human Tissues: Towards the Inosinome Atlas." *Scientific Reports* 5 (October): 14941.

Picelli, Simone, Åsa K. Björklund, Omid R. Faridani, Sven Sagasser, Gösta Winberg, and Rickard Sandberg. 2013. "Smart-seq2 for Sensitive Full-Length Transcriptome Profiling in Single Cells." *Nature Methods* 10 (11): 1096–98.

Porath, Hagit T., Shai Carmi, and Erez Y. Levanon. 2014. "A Genome-Wide Map of Hyper-Edited RNA Reveals Numerous New Sites." *Nature Communications* 5 (August): 4726.

Prasanth, Kannanganattu V., Supriya G. Prasanth, Zhenyu Xuan, Stephen Hearn, Susan M. Freier, C. Frank Bennett, Michael Q. Zhang, and David L. Spector. 2005. "Regulating Gene Expression through RNA Nuclear Retention." *Cell* 123 (2): 249–63.

Quesada, Víctor, Laura Conde, Neus Villamor, Gonzalo R. Ordóñez, Pedro Jares, Laia Bassaganyas, Andrew J. Ramsay, et al. 2011. "Exome Sequencing Identifies Recurrent Mutations of the Splicing Factor SF3B1 Gene in Chronic Lymphocytic Leukemia." *Nature Genetics* 44 (1): 47–52.

Ramaswami, Gokul, and Jin Billy Li. 2014. "RADAR: A Rigorously Annotated Database of A-to-I RNA Editing." *Nucleic Acids Research* 42 (Database issue): D109–13.

Rand, Arthur C., Miten Jain, Jordan M. Eizenga, Audrey Musselman-Brown, Hugh E. Olsen, Mark Akeson, and Benedict Paten. 2017. "Mapping DNA Methylation with High-Throughput Nanopore Sequencing." *Nature Methods* 14 (4): 411–13.

Rice, Gillian I., Paul R. Kasher, Gabriella M. A. Forte, Niamh M. Mannion, Sam M. Greenwood, Marcin Szynkiewicz, Jonathan E. Dickerson, et al. 2012. "Mutations in ADAR1 Cause Aicardi-

Goutières Syndrome Associated with a Type I Interferon Signature." *Nature Genetics* 44 (11): 1243–48.

Rivas, Manuel A., Matti Pirinen, Donald F. Conrad, Monkol Lek, Emily K. Tsang, Konrad J. Karczewski, Julian B. Maller, et al. 2015. "Human Genomics. Effect of Predicted Protein-Truncating Genetic Variants on the Human Transcriptome." *Science* 348 (6235): 666–69.

Robinson, Eva, and Alison Tang. 2020. "Brooks Lab Western Blotting Protocol." Protocols.io. March 20, 2020. https://doi.org/10.17504/protocols.io.bcsmiwc6.

Rossi, Davide, Alessio Bruscaggin, Valeria Spina, Silvia Rasi, Hossein Khiabanian, Monica Messina, Marco Fangazio, et al. 2011. "Mutations of the SF3B1 Splicing Factor in Chronic Lymphocytic Leukemia: Association with Progression and Fludarabine-Refractoriness." *Blood* 118 (26): 6904–8.

Roundtree, Ian A., Molly E. Evans, Tao Pan, and Chuan He. 2017. "Dynamic RNA Modifications in Gene Expression Regulation." *Cell* 169 (7): 1187–1200.

Rozowsky, Joel, Alexej Abyzov, Jing Wang, Pedro Alves, Debasish Raha, Arif Harmanci, Jing Leng, et al. 2011. "AlleleSeq: Analysis of Allele-Specific Expression and Binding in a Network Framework." *Molecular Systems Biology* 7 (August): 522.

Rueter, S. M., T. R. Dawson, and R. B. Emeson. 1999. "Regulation of Alternative Splicing by RNA Editing." *Nature* 399 (6731): 75–80.

Saiki, R. K., D. H. Gelfand, S. Stoffel, S. J. Scharf, R. Higuchi, G. T. Horn, K. B. Mullis, and H. A. Erlich. 1988. "Primer-Directed Enzymatic Amplification of DNA with a Thermostable DNA Polymerase." *Science* 239 (4839): 487–91.

Scadden, A. D. J. 2005. "The RISC Subunit Tudor-SN Binds to Hyper-Edited Double-Stranded RNA and Promotes Its Cleavage." *Nature Structural & Molecular Biology* 12 (6): 489–96.

———. 2007. "Inosine-Containing dsRNA Binds a Stress-Granule-like Complex and Downregulates Gene Expression in Trans." *Molecular Cell* 28 (3): 491–500.

Schurch, Nicholas J., Pietá Schofield, Marek Gierliński, Christian Cole, Alexander Sherstnev, Vijender Singh, Nicola Wrobel, et al. 2016. “How Many Biological Replicates Are Needed in an RNA-Seq Experiment and Which Differential Expression Tool Should You Use?” *RNA* 22 (6): 839–51.

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

Shafin, Kishwar, Trevor Pesout, Pi-Chuan Chang, Maria Nattestad, Alexey Kolesnikov, Sidharth Goel, Gunjan Baid, et al. 2021. “Haplotype-Aware Variant Calling with PEPPER-Margin-DeepVariant Enables High Accuracy in Nanopore Long-Reads.” *Nature Methods* 18 (11): 1322–32.

Sharon, Donald, Hagen Tilgner, Fabian Grubert, and Michael Snyder. 2013. “A Single-Molecule Long-Read Survey of the Human Transcriptome.” *Nature Biotechnology* 31 (11): 1009–14.

Sibbritt, Tennille, Hardip R. Patel, and Thomas Preiss. 2013. “Mapping and Significance of the mRNA Methylome.” *Wiley Interdisciplinary Reviews. RNA* 4 (4): 397–422.

Simpson, Jared T., Rachael E. Workman, P. C. Zuzarte, Matei David, L. J. Dursi, and Winston Timp. 2017. “Detecting DNA Cytosine Methylation Using Nanopore Sequencing.” *Nature Methods* 14 (4): 407–10.

Smith, Andrew M., Miten Jain, Logan Mulroney, Daniel R. Garalde, and Mark Akeson. 2019a. “Reading Canonical and Modified Nucleobases in 16S Ribosomal RNA Using Nanopore Native RNA Sequencing.” *PLOS ONE*. https://doi.org/10.1371/journal.pone.0216709.

———. 2019b. “Reading Canonical and Modified Nucleobases in 16S Ribosomal RNA Using Nanopore Native RNA Sequencing.” *PloS One* 14 (5): e0216709.

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.

Sommer, B., M. Köhler, R. Sprengel, and P. H. Seeburg. 1991. "RNA Editing in Brain Controls a Determinant of Ion Flow in Glutamate-Gated Channels." *Cell* 67 (1): 11–19.

Soneson, Charlotte, Yao Yao, Anna Bratus-Neuenschwander, Andrea Patrignani, Mark D. Robinson, and Shobbir Hussain. n.d. "A Comprehensive Examination of Nanopore Native RNA Sequencing for Characterization of Complex Transcriptomes." https://doi.org/10.1101/574525.

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.

Subtelny, Alexander O., Stephen W. Eichhorn, Grace R. Chen, Hazel Sive, and David P. Bartel. 2014. "Poly(A)-Tail Profiling Reveals an Embryonic Switch in Translational Control." *Nature* 508 (7494): 66–71.

Sultan, Marc, Marcel H. Schulz, Hugues Richard, Alon Magen, Andreas Klingenhoff, Matthias Scherf, Martin Seifert, et al. 2008. "A Global View of Gene Activity and Alternative Splicing by Deep Sequencing of the Human Transcriptome." *Science* 321 (5891): 956–60.

Sun, Shuying, Zuo Zhang, Rahul Sinha, Rotem Karni, and Adrian R. Krainer. 2010. "SF2/ASF Autoregulation Involves Multiple Layers of Post-Transcriptional and Translational Control." *Nature Structural & Molecular Biology* 17 (3): 306–12.

Tajaddod, Mansoureh, Michael F. Jantsch, and Konstantin Licht. 2016. "The Dynamic Epitranscriptome: A to I Editing Modulates Genetic Information." *Chromosoma* 125 (1): 51–63.

Tang, Alison D., Cameron M. Soulette, Marijke J. van Baren, Kevyn Hart, Eva Hrabeta-Robinson, Catherine J. Wu, and Angela N. Brooks. 2018. "Full-Length Transcript Characterization of

SF3B1 Mutation in Chronic Lymphocytic Leukemia Reveals Downregulation of Retained Introns." *bioRxiv*. https://doi.org/10.1101/410183.

Tang, Sze Jing, Haoqing Shen, Omer An, Huiqi Hong, Jia Li, Yangyang Song, Jian Han, et al. 2020. "Cis- and Trans-Regulations of Pre-mRNA Splicing by RNA Editing Enzymes Influence Cancer Development." *Nature Communications* 11 (1): 799.

Tardaguila, Manuel, Lorena de la Fuente, Cristina Marti, Cécile Pereira, Francisco Jose Pardo-Palacios, Hector Del Risco, Marc Ferrell, et al. 2018. "SQANTI: Extensive Characterization of Long-Read Transcript Sequences for Quality Control in Full-Length Transcriptome Identification and Quantification." *Genome Research*, February. https://doi.org/10.1101/gr.222976.117.

Tavakoli, Sepideh, Mohammad Nabizadehmashhadtoroghi, Amr Makhamreh, Howard Gamper, Neda K. Rezapour, Ya-Ming Hou, Meni Wanunu, and Sara H. Rouhanifard. 2021. "Detection of Pseudouridine Modifications and Type I/II Hypermodifications in Human mRNAs Using Direct, Long-Read Sequencing." *bioRxiv*. https://doi.org/10.1101/2021.11.03.467190.

Temin, H. M., and S. Mizutani. 1970. "RNA-Dependent DNA Polymerase in Virions of Rous Sarcoma Virus." *Nature* 226 (5252): 1211–13.

Ten Hacken, Elisa, Rebecca Valentin, Fara Faye D. Regis, Jing Sun, Shanye Yin, Lillian Werner, Jing Deng, et al. 2018. "Splicing Modulation Sensitizes Chronic Lymphocytic Leukemia Cells to Venetoclax by Remodeling Mitochondrial Apoptotic Dependencies." *JCI Insight* 3 (19). https://doi.org/10.1172/jci.insight.121438.

Tilgner, Hagen, Fabian Grubert, Donald Sharon, and Michael P. Snyder. 2014. "Defining a Personal, Allele-Specific, and Single-Molecule Long-Read Transcriptome." *Proceedings of the National Academy of Sciences of the United States of America* 111 (27): 9869–74.

Venturini, Luca, Shabhonam Caim, Gemy George Kaithakottil, Daniel Lee Mapleson, and David Swarbreck. 2018. "Leveraging Multiple Transcriptome Assembly Methods for Improved Gene Structure Annotation." *GigaScience* 7 (8). https://doi.org/10.1093/gigascience/giy093.

Volden, Roger, Theron Palmer, Ashley Byrne, Charles Cole, Robert J. Schmitz, Richard E. Green, and Christopher Vollmers. 2018. “Improving Nanopore Read Accuracy with the R2C2 Method Enables the Sequencing of Highly Multiplexed Full-Length Single-Cell cDNA.” *Proceedings of the National Academy of Sciences of the United States of America* 115 (39): 9726–31.

Vollger, Mitchell R., Glennis A. Logsdon, Peter A. Audano, Arvis Sulovari, David Porubsky, Paul Peluso, Gregory T. Concepcion, et al. 2019. “Improved Assembly and Variant Detection of a Haploid Human Genome Using Single-Molecule, High-Fidelity Long Reads.” *bioRxiv*. https://doi.org/10.1101/635037.

Vollmers, Apple Cortez, Honey E. Mekonen, Sophia Campos, Susan Carpenter, and Christopher Vollmers. 2021. “Generation of an Isoform-Level Transcriptome Atlas of Macrophage Activation.” *The Journal of Biological Chemistry* 296 (January): 100784.

Wang, Eric T., Rickard Sandberg, Shujun Luo, Irina Khrebtukova, Lu Zhang, Christine Mayr, Stephen F. Kingsmore, Gary P. Schroth, and Christopher B. Burge. 2008. “Alternative Isoform Regulation in Human Tissue Transcriptomes.” *Nature* 456 (7221): 470–76.

Wang, Lili, Angela N. Brooks, Jean Fan, Youzhong Wan, Rutendo Gambe, Shuqiang Li, Sarah Hergert, et al. 2016. “Transcriptomic Characterization of SF3B1 Mutation Reveals Its Pleiotropic Effects in Chronic Lymphocytic Leukemia.” *Cancer Cell* 30 (5): 750–63.

Wang, Lili, Michael S. Lawrence, Youzhong Wan, Petar Stojanov, Carrie Sougnez, Kristen Stevenson, Lillian Werner, et al. 2011. “SF3B1 and Other Novel Cancer Genes in Chronic Lymphocytic Leukemia.” *The New England Journal of Medicine* 365 (26): 2497–2506.

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.

Wan, Youzhong, and Catherine J. Wu. 2013. “SF3B1 Mutations in Chronic Lymphocytic Leukemia.” *Blood* 121 (23): 4627–34.

Weirather, Jason L., Mariateresa de Cesare, Yunhao Wang, Paolo Piazza, Vittorio Sebastiano, Xiu-Jie Wang, David Buck, and Kin Fai Au. 2017. “Comprehensive Comparison of Pacific Biosciences and Oxford Nanopore Technologies and Their Applications to Transcriptome Analysis.” *F1000Research* 6 (February): 100.

Will, C. L., C. Schneider, A. M. MacMillan, N. F. Katopodis, G. Neubauer, M. Wilm, R. Lührmann, and C. C. Query. 2001. “A Novel U2 and U11/U12 snRNP Protein That Associates with the Pre-mRNA Branch Site.” *The EMBO Journal* 20 (16): 4536–46.

Workman, Rachael E., Alison Tang, Paul S. Tang, Miten Jain, John R. Tyson, Philip C. Zuzarte, Timothy Gilpatrick, et al. 2018. “Nanopore Native RNA Sequencing of a Human poly(A) Transcriptome.” *bioRxiv*. https://doi.org/10.1101/459529.

Wurster, Claudia D., and Albert C. Ludolph. 2018. “Nusinersen for Spinal Muscular Atrophy.” *Therapeutic Advances in Neurological Disorders* 11 (March): 1756285618754459.

Wu, Thomas D., and Colin K. Watanabe. 2005. “GMAP: A Genomic Mapping and Alignment Program for mRNA and EST Sequences.” *Bioinformatics* 21 (9): 1859–75.

Yi, Hyerim, Joha Park, Minju Ha, Jaechul Lim, Hyeshik Chang, and V. Narry Kim. 2018. “PABP Cooperates with the CCR4-NOT Complex to Promote mRNA Deadenylation and Block Precocious Decay.” *Molecular Cell* 70 (6): 1081–88.e5.

Yin, Shanye, Rutendo G. Gambe, Jing Sun, Aina Zurita Martinez, Zachary J. Cartun, Fara Faye D. Regis, Youzhong Wan, et al. 2019. “A Murine Model of Chronic Lymphocytic Leukemia Based on B Cell-Restricted Expression of Sf3b1 Mutation and Atm Deletion.” *Cancer Cell* 35 (2): 283–96.e5.

Young, Matthew D., Matthew J. Wakefield, Gordon K. Smyth, and Alicia Oshlack. 2012. “Goseq: Gene Ontology Testing for RNA-Seq Datasets.” *R Bioconductor*. https://bioconductor.org/packages/devel/bioc/vignettes/goseq/inst/doc/goseq.pdf.