Full-length transcript characterization of SF3B1 mutation in chronic lymphocytic leukemia reveals downregulation of retained introns

Alison D. Tang 1, Cameron M. Soulette 2, Marijke J. van Baren 1, Kevyn Hart 1, Eva Hrabeta-Robinson 1, Catherine J. Wu 3,4,5 & Angela N. Brooks 1✉

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 downregulation 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.
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)1–4, uveal melanoma5–7, breast cancer8–9, and myelodysplastic syndromes10,11. 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-mRNA12–15. In CLL, mutations in the HEAT-repeat domain of SF3B1, such as the K700E hotspot mutation, have been shown to associate with poor clinical outcome1–4. 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 SF3B116. In addition, 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′ splicing17–19. Mutant SF3B1–associated changes in branch point recognition and usage20–22 form the model in which mutant SF3B1 affects acceptor splice sites. Targeting a mutant SF3B1-induced mis-spliced exon in the tumor suppressor BRD9 using antisense oligonucleotides suppresses tumor growth23, further revealing the therapeutic implications for treating mutant SF3B1.

While alternative splicing (AS) patterns induced by SF3B1 mutations have been examined on an event level with short reads, these patterns have not been systematically examined on an isoform level. Studying splicing factor mutations at the level of single splice junctions limits our complete understanding of the functional consequences of these aberrant splicing changes. Short reads provide an incomplete picture of the transcriptome because exon connectivity is difficult to determine from fragments24. Moreover, detection and quantification of transcripts containing retained introns using short reads is difficult and often overlooked25. Long-read-sequencing techniques now offer increased information on exon connectivity by sequencing full-length transcript molecules26–28. Nanopore sequencing works by measuring the change in electrical current caused by DNA or RNA threading through a nanopore and converting the signal into nucleotide sequences29,30. Nanopore sequencing yields reads as long as 2 megabases31 and has been used for applications ranging from the sequencing of the centromere of the Y chromosome32, the human genome33, and single-cell transcriptome sequencing34. However, nanopore technology has yet to be thoroughly explored as a tool for differential splicing, nor has it been used to explore cancer-associated splicing factor mutations. The high-throughput Oxford Nanopore PromethION has the potential to yield upward of 100 gigabases (Gb) per flow cell35; the substantial number of long molecules sequenced makes the PromethION applicable to our purposes of transcript detection and quantification.

To investigate SF3B1K700E 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 ~85–87%36–38, although accuracy can change depending on iterations of the technology and library preparation methods39. In contrast, the long consensus reads produced by Pacific Biosciences are able to achieve much higher base accuracies39. Thus, software tools developed for PacBio sequencing were not developed for noisier data and may disregard raw reads with less than 99% accuracy40 or discard assembled isoforms with too many errors41. To assemble isoforms and perform splicing analysis from nanopore reads, we have created a workflow called Full-Length Alternative Isoform analysis of RNA (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-Seq17, 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 control42. Following the identification of high-confidence isoforms from nanopore data, FLAIR provides a framework for performing AS and differential isoform usage analyses. Upon splicing analysis of the nanopore CLL data, we observe a bias toward increased alternative 3′ splice sites over alternative 5′ splice sites in CLL SF3B1K700E samples, consistent with the known effects of SF3B1 mutation. We also highlight a previously underappreciated finding of differential intron retention in CLL SF3B1K700E versus SF3B1WT 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 biological insights into SF3B1 mutations and demonstrates the potential for discovering cancer biology with long-read sequencing.

**Results**

**FLAIR identifies full-length isoforms in CLL**. As a pilot study to characterize SF3B1K700E full-length transcripts, we performed nanopore 2D cDNA minION sequencing of a CLL SF3B1K700E sample that was cloned for the mutation and a CLL SF3B1WT sample that did not contain other common CLL-associated mutations. These samples were previously examined in a study of AS of SF3B1 mutants using short-read RNA-Seq data17. With the pilot nanopore data, we found an increase in aberrant 3′ splicing in CLL SF3B1K700E (Supplementary Fig. 1), recapitulating findings from the CLL short-read data17. This motivated the resequencing of additional replicate samples with nanopore 1D cDNA sequencing on the PromethION, a high-throughput nanopore sequencer (Fig. 1a). From the same cohort, we sequenced 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”). Although the RNA samples had been stored for 5 years, they showed minimal signs of degradation as the read lengths were comparable to those observed from other nanopore cDNA-sequencing runs26,38,43. We observed relatively strong gene expression correlation (average Pearson correlation 0.822) between the three samples sequenced on both the minION and promethION (Supplementary Fig. 2), and so combined the data for subsequent analyses.

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. 1b). 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 minimap244 and GMAP45, the latter of
indels, small gaps in the read alignments were artificially filled (“Methods”). FLAIR only considers correct splice sites as those from a curated set of splice junctions, such as splice sites present in annotations. 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. Given that SF3B1 mutations can subtly alter 3′ splice-site choice\textsuperscript{17,19,22} and accurate representation of splice-site usage is crucial, we increased splice-site accuracy by correcting read alignments using splice sites in either annotations or matched CLL short-read sequencing (Fig. 2a, Supplementary Fig. 3c, d). 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. 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 splice junction linkage by constraining reads to align only to splice junction chains with additional support. The realignment was also crucial for better distinguishing splice-site differences\textsuperscript{17}. Finally, FLAIR filters out first-pass isoforms that have fewer than three supporting reads and the remaining isoforms with sufficient coverage comprise the final high-confidence nanopore isoform set. Using FLAIR, we identified a total of 326,699 high-confidence spliced isoforms (Supplementary Data 1, 2). 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 (Supplementary Fig. 4). For all conditions, we found that sequencing at greater depth would facilitate the discovery of even more isoforms. At all read depths, we were able to detect the greatest number of isoforms in CLL SF3B1\textsuperscript{K700E}. The read lengths of CLL SF3B1\textsuperscript{WT} were noticeably shorter, resulting in fewer isoforms able to be detected compared with B cell or CLL SF3B1\textsuperscript{K700E}. The saturation analysis illustrates the diversity of isoforms resulting from mutant SF3B1 as well as the importance of read length for isoform discovery.

**FLAIR improves isoform detection from nanopore reads.** We evaluated FLAIR against other nanopore isoform assembly methods: Mandalorion\textsuperscript{34} and Oxford Nanopore’s Pinfish
We simulated a set of nanopore reads using minimap2 and selecting alignments using either a transcriptome-alignment approach or its alignment-based mode with all minimap2 alignments, as designed for 2D reads\(^{34}\), however, has been reworked to permit isoform building accurately. We benchmarked using GFFCompare\(^{48}\) on assemblies using both a simulated dataset and Lexogen Spike-In RNA Variants (SIRV)\(^{37}\) sequenced with nanopore 1D technology. We simulated a set of nanopore transcriptome reads by creating a wrapper for NanoSim\(^{49}\), a tool created for simulating genomic nanopore reads ("Methods"). The SIRV RNAs are of known sequence and concentration and are of comparable complexity to human transcripts\(^{50}\). We did not benchmark FLAIR with short-read data as Mandalorion and Pinfish do not incorporate short reads into their algorithms for a splice-site correction step. Instead, Mandalorion uses raco\(^{31}\) to generate consensus isoform sequences from clustered reads. Mandalorion was initially designed for 2D reads\(^{34}\), however, has been reworked to perform best on higher-accuracy R2C2 reads\(^{37}\). Pinfish clusters nanopore reads and generates consensus isoforms using the median exon boundaries from all transcripts in the cluster. We observed that FLAIR had higher sensitivity than other methods and had comparable or higher precision (Fig. 2b and Supplementary Fig. 5). Based on the base-level and intron chain-level sensitivities, Pinfish appeared to require more reads in order to determine isoform structures. We also noted that as the supporting read threshold was increased, there was an expected decrease in sensitivity and negligible increase in precision. As a result, we decided to use FLAIR with a threshold of 3, increasing the number of isoforms assembled to better capture all splicing alterations.

Quantifying expression of isoforms using nanopore data may appear straightforward in that each nanopore read corresponds to a single transcript present in the cell. However, ambiguity of read assignments to transcripts can occur when assigning noisy, truncated reads to genes with greater isoform complexity. We benchmarked different approaches of assigning reads to transcripts using the 1D SIRV reads and the simulated nanopore reads using either a transcriptome-alignment approach or salmon\(^{52}\). The former involved aligning reads to transcript sequences using minisnap2 and selecting alignments using either primary read alignments or read alignments with mapping quality scores greater than or equal to 1. The salmon algorithm is able to account for sequence-specific, fragment-GC, and positional biases. We evaluated salmon as a quantification tool using its alignment-based mode with all minisnap2 alignments, as performance of the quasi-alignment mode could suffer from low read accuracy. We observed that counting transcript alignments with MAPQ \(\geq 1\) correlated the best with expected counts for both the SIRV and simulated data (Fig. 2c) and used this quantification approach for our analyses. Although the correlations using salmon were not as strong as the MAPQ \(\geq 1\) approach, the latter approach filters out many low-quality reads and is unable to include them in the counts; in the SIRVs, 25.5% (5,469 out of 21,456) of reads have MAPQ of 0 and were discarded from quantification. We recognized the increased inclusion of reads from the data when using salmon and provided an option to use salmon with FLAIR.

### Nanopore read analysis detects 3' splice-site changes

Previous studies have demonstrated that SF3B1\(^{K700E}\) promotes alternative 3' splice-site (3'SS) usage\(^{17}\), a pattern we sought to validate in our nanopore data. From a cohort of 37 CLL samples\(^{17}\) (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 junkBASE\(^{53}\). 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. 3a). 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; we therefore wanted to identify the altered 3' splicing events that would be significantly altered by mutant SF3B1 in the nanopore data.

Due to the sheer complexity of alternatively spliced isoforms, we wrote an AS event caller for FLAIR to more accurately classify AS in isoforms ("Methods"). 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. Using DRIMSeq\(^{54}\) for statistical testing, 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\(^{WT}\) (Supplementary Data 3). 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.

### Table 1 Nanopore-sequencing statistics

| Sample          | Data (Gb) | Total reads | % Pass | % full length | Median pass read length | Mean pass read length | No. of genes detected | % genes with AS |
|-----------------|-----------|-------------|--------|---------------|-------------------------|-----------------------|----------------------|----------------|
| Promethion WT 1 | 2.75      | 5,839,921   | 23.1   | 22.9          | 816                     | 1020                  | 17,369               | 53.0%          |
| Promethion WT 2 | 0.526     | 2,051,438   | 6.91   | 27.7          | 768                     | 936                   | 8492                 | 43.9%          |
| Promethion WT 3 | 52.2      | 63,378,471  | 56.9   | 28.0          | 788                     | 944                   | 43,223               | 41.7%          |
| Promethion MT 1 | 0.522     | 3,016,571   | 1.70   | 24.0          | 718                     | 962                   | 6272                 | 31.8%          |
| Promethion MT 2 | 91.1      | 93,130,856  | 74.8   | 30.2          | 870                     | 1020                  | 45,744               | 41.2%          |
| Promethion MT 3 | 3.1       | 15,101,288  | 5.38   | 30.2          | 919                     | 1123                  | 16,218               | 54.7%          |
| Promethion B cell 1 | 65.8 | 66,311,147  | 60.7   | 34.3          | 948                     | 1099                  | 33,538               | 46.6%          |
| Promethion B cell 2 | 0.970 | 5,390,853   | 7.16   | 33.5          | 712                     | 951.3                 | 13,638               | 44.0%          |
| Promethion B cell 3 | 0.731 | 2,934,130   | 4.80   | 27.8          | 766                     | 1015                  | 9372                 | 34.6%          |
| Minion WT 1     | 0.105     | 119,091     | -      | 19.6          | 651                     | 866                   | 8320                 | 36.6%          |
| Minion WT 2     | 0.138     | 129,798     | -      | 25.5          | 880                     | 1084                  | 9747                 | 42.6%          |
| Minion WT 3     | 0.042     | 46,077      | -      | 32.5          | 749                     | 922                   | 5379                 | 26.8%          |

For each run, the following statistics were computed: gigabase of data, total reads, the percentage of reads designated as passing by the basecaller for PromethION runs, percentage of aligning pass reads that are full-length, median pass read length, mean pass read length, the number of genes covered, and the number of genes with evidence of alternative splicing ("Methods"). For the 2D MinION runs, the total number of reads is the number of 2D reads, which was also considered the total number of pass reads.

NATURE COMMUNICATIONS | https://doi.org/10.1038/s41467-020-15171-6

ARTICLE NATURE COMMUNICATIONS | https://doi.org/10.1038/s41467-020-15171-6

www.nature.com/naturecommunications
reads18,19. From the nanopore-sequencing analysis, the distribution of distances between splice sites determined using the GMAP (yellow) or minimap2 (blue and green) spliced aligners to the closest annotated splice sites for a subset of the CLL SF3B1K700E reads; the minimap2 aligned vs annotated splice-site histogram was further distinguished between alignments that had (green) or had not (blue) been subject to FLAIR splice correction. Base-level sensitivity (left) and precision (right) of isoform sets built using Mandalorion (green), Pinfish (yellow), or FLAIR (blue). Each software tool was run with varying supporting read thresholds (<bottom row>). First column: counting primary alignments, second column: salmon given all alignments, third column: counting MAPQ ≥ 1 alignments. Boxplots show median line, box limits are upper and lower quartile, and whiskers are 1.5× interquartile.

We were unable to find any unifying sequence motif associated with these altered 3′SSs identified in the nanopore data. 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 (Supplementary Fig. 6a). This motif is concordant with other mutant SF3B1 studies using short reads18,19. From the nanopore-sequencing analysis, the distribution of distances between SF3B1K700E-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 × 10−2) (Fig. 3b), similar to what has been reported in CLL short-read sequencing17. One of the alternative 3′SSs identified from both long and short reads was in the ERGIC3 gene (Fig. 3c). There were two dominant isoforms: a novel isoform containing the proximal splice site that was more abundant in SF3B1K700E 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′SSs 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.

**SF3B1** downregulates intron retention events. Intron retentions (IR) have been observed to differentiate tumors from matched normal tissue, as they are highly prevalent across a variety of cancers. However, based on common approaches used by short-read AS analysis tools, it is difficult to characterize IR event usage confidently using short reads. Thus, unless stringent approaches are used, intron retention events are easily misclassified particularly in regions with complex AS. 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**, 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** sample compared with CLL **SF3B1**WT (Fig. 4a). When performing the same comparison between B cell and **SF3B1**WT, 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. 4b).

To further investigate the effect of **SF3B1** on increased intron splicing, we reanalyzed Nalm-6 Pre-B isogenic cell lines with either **SF3B1** WT or **SF3B1**K700E sequenced using short reads. We used juncBASE 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 (Supplementary Fig. 7a), 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 (Supplementary Fig. 7b). 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 (Supplementary Fig. 7c).

Seeing that the trend of higher IR expression in **SF3B1**WT was observed transcriptome-wide in the nanopore data, we narrowed...
Fig. 4 Intron retention events are more strongly downregulated in CLL SF3B1K700E. **a** Expression fold-change between SF3B1K700E and SF3B1WT of FLAIR isoforms with (IR) or without (spliced) retained introns. Boxplot median difference = 0.395. Source data are provided in a Source Data file. **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. Source data are provided in a Source Data file. **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 color 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).

We then looked at splice-site motifs for the differentially retained introns (Fig. 4f and Supplementary Fig. 6). 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/TGAG motif 15 bp upstream of the 3'SS (Fig. 4f). 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.22 Sequence analysis of introns with increased inclusion in SF3B1WT 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 (Supplementary Fig. 7d, e), although not at exactly the same position. This further supports an underlying mechanism of SF3B1K700E in which the mutant prefers splicing at a 3'SS ~15 bp downstream of a strong branch point22. We did not observe the same motifs for the IR events identified from our focus to only the introns that were significantly differentially retained between SF3B1K700E and SF3B1WT. Using DRIMSeq 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 short-read-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. 4c, Supplementary Data 3). An example of a gene with increased expression of an isoform with increased IR splicing in CLL SF3B1K700E is ADTRP (androgen dependent TFPI regulating protein), which is involved with blood coagulation. Long reads enabled the identification of coordinated splicing events in ADTRP, such as isoforms with a differentially skipped exon coordinated with the differentially retained intron (Fig. 4d, e).
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 (Supplementary Fig. 8). 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.

Taking full advantage of the long-read data and using our assembled transcripts and isoform quantification approaches, we identified 143 full-length isoforms that were differentially used (corrected p < 0.1 and abs(dPSI) > 10) between CLL SF3B1K700E and SF3B1WT (Supplementary Data 4). Of these differentially used isoforms, 53 isoforms contained AS events that were significantly different between mutant and wild type SF3B1. Collectively, these isoforms contained 5 alternative 3′SSs, 1 alternative 5′SS, eight retained introns, and 28 exon skipping events that were tested to be significant. Although only a few differentially used isoforms contained more than one significant splicing alteration, nearly all (95.1%) overlapped two or more regions where FLAIR identified multiple splicing choices. Differentially used isoforms in the LINC01089, LINC01480, regions where FLAIR identified multiple splicing choices. Alternatively used isoforms in the LINC01089, LINC01480, XBP1 genes contain more IR or partial IR in SF3B1WT relative to SF3B1K700E (Supplementary Fig. 9a–c). These alternative isoforms with retained introns were also coupled with alternative 3′ splicing events, which is a type of regulated coupling that would be difficult to determine confidently with short-read data.

**SF3B1K700E 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)\(^{19}\). 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\(^{60,61}\) (Fig. 5a). 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\(^ {61-64}\). For example, SRSF1 has several unproductive transcripts that are known to be either nuclear-retained or NMD-triggering\(^ {62}\), two of which (ENST00000581979.5 and unannotated Isoform V\(^ {62}\)) were able to identify and accurately predict as unproductive in our nanopore data (Supplementary Fig. 10). We also identified 5 additional unannotated SRSF1 isoforms with more than 100 supporting reads, 2 of which are productive and 3 are unproductive.

We found that isoform productivity predicted from local alterations often matched the productivity of the full isoform. That is, for the alternative 3′SSs with more proximal site usage in SF3B1K700E, the alterations that were predicted to preserve the reading frame (i.e., those that included a multiple of three basepairs of sequence in the coding region) were more associated with productive isoforms (Table 2). However, there were some exceptions in which alterations that were predicted to disrupt the reading frame were associated with productive isoforms and vice versa, demonstrating the need for long-read data to accurately predict productivity changes caused by alterations at a junction level. Next, we looked systemically at all FLAIR isoforms categorized by their productivity in combination with the presence or absence of retained introns (Fig. 5b). Although together productive and unproductive IR isoforms were downregulated in SF3B1K700E (Fig. 4a), the reduction was more pronounced in the unproductive IR isoforms (productive-spliced and unproductive-IR two-sided Mann–Whitney U p = 1.25 × 10\(^ {-6}\)). To further understand the decrease in unproductive IR observed in SF3B1K700E, 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 (Supplementary Fig. 11 and Supplementary Data 5). The prevalence of cellular signaling GO terms parallels a finding in glioblastoma, where genes with a decrease in retained introns regulated by PRMT5 are also associated with perturbed kinase signaling\(^ {65}\).

**Discussion**

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 AS 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 net 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 SF3B1K700E. The alternative 3′SS usage patterns were consistent with alterations identified in short-read data. In addition, long-read sequencing highlighted a downregulation of isoforms containing retained introns in SF3B1K700E relative to SF3B1WT. This downregulation was corroborated by reanalyzing CLL, Nalm-6 cell line, 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\(^ {16,66}\). The subset of introns that exhibited increased splicing in the mutant point to a different intron retention landscape in CLL SF3B1K700E. 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\(^ {62}\).

Full-length reads also allow for improved identification of IR 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\(^ {19}\). 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 GO analyses revealed kinase signaling associated with the unproductive IR events downregulated in SF3B1K700E. We postulate that these unproductive-retained introns are the cases of detained introns, as kinase signaling has been associated specifically with detained introns65. The 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 AS 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 (<1000 bp) sequenced in long reads (Supplementary Fig. 8). Everything considered, studying splicing factor mutations in primary 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 samples per genotype being the minimum. Even though short-read technology is able to sequence more deeply than long reads, the ability of short reads to saturate splice junction detection is depth-dependent, thus splicing study results will not be as deep as possible. Fortunately, the throughput and accuracy of nanopore and PacBio technology has the potential to increase with subsequent iterations of the technologies. For nanopore in particular, methods to achieve higher sequence accuracy or circumvent PCR bias and reverse transcription length restrictions have been developed. In line with the rigorous pace of improvements in the field of long reads, PacBio has recently improved their throughput 8× with the newest PacBio Sequel II system, which has been shown to generate ~19 and 83 Gb of consensus reads. 

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.

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 or normal samples using the Qiagen RNeasy Mini Kit17. The sample IDs of the CLL SF3B1WT samples are CW67 (WT 1), CW95 (WT 3), and JGG0035 (WT 2) and the IDs of the SF3B1WT samples are DFCI-5067 (MT 1), CL043/CW109 (MT 2), and CL032/CW84 (MT 3) from Wang et al.17. GTEX was not included in the RNAseq and RNBio libraries above 7. The extracted RNA was reverse transcribed using the SmartSeq protocol1 and cDNA was PCR-amplified using KAPA HiFi ReadyMix 2× (98 °C for 20 s, 67 °C for 16 s, 72 °C for 4 min)13. Fifteen cycles of PCR were performed. Prior to library preparation, the concentration of the cDNA for the samples ranged from 1.26 to 10.7 ng/µl (Supplementary Table 1). Oxford MinION 2D amplicon libraries were generated according to the Nanopore community protocol using library preparation kit SQK-LSK208 and sequenced on R9 flowcells. Basescalling was performed with albacore v1.1.0 2D basescalling using the—flow cell FLO-MIN107 and—kkit 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 0.8× 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 (Supplementary Table 2). Basescalling of 1D PromethION reads was done with guppy v3.3.5 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. (1) adapters are aligned to all the reads using blastn; (2) if there are at least ten 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 bwa mem to the GENCODE v24 transcriptome and the read-isoform assignments were determined using the primarily 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 hg38 (http://hgdownload.cse.ucsc.edu/hg38/) using bwa-mem to the GENCODE v24 comprehensive annotations or short reads. Splice junctions were extracted from matched short-read data, and only the junctions supported by three 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.

**Data simulation.** We created a wrapper script for NanoSim to generate simulated transcriptomic data. We simulated reads from transcripts of genes that were found to be expressed in our real nanopore data, with read lengths modeled after real data using our wrapper script (see “Code availability”). We used the hg38 nanopore error model simulated from step 1 of NanoSim.

**Isoform-identification methods.** We used Mandalorion version II and FLAIR v1.4 to assemble isoforms on the SRV and simulated data with tuned read support threshold parameters. FLAIR was run with the default settings. Mandalorion was run with the parameters on the GitHub—u -s -d -s -z -200 -r 0.05 -i 0 -l 100 -t -T 60 varying -R (minimum number of reads for an isoform to be reported). Consensus isoform sequences were aligned to the genome with minimap2 and converted to gtf prior to running GFFCompare. Pinush was run with default parameters, adjusting the minimum cluster size parameter only. 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 (> 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 MAPQ ≥ 1. Three supporting reads with MAPQ ≥ 10 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 three 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.

**AS 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’SSs were grouped by the 5’ SS they were observed with and had to be present in overlapping exons and vice versa. The alternative donor junctions were defined as donor junctions 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 A5 events. For analysis of the PromethION data with replicates, we used DRIMSeq\(^1\). DRIMSeq statistical testing accounted for sequencing and RNA batch according to batch numbers for each sample (Supplementary Table 1). 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 should be from either the CLL SF3B1 WT or SF3B1 K70E condition. A pseudocount of 1 was used for events with 1–2 dropout samples from being excluded from testing. For differential isoform usage testing (FLAIR diffExpr), isoforms were grouped by gene and only genes with at least 25 reads in 4 of the 6 samples were tested. As these libraries were poly(A) selected, we did not distinguish between IR due to incomplete transcript processing and IR 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 ten or more in one of the conditions were plotted. IR were defined as an 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 site\(^6\). 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 v1.3.2\(^6\), 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 supplementary Table 1). The expression docount of 1 was used to prevent events with 1 dropout samples from being excluded from testing. For analysis of the PromethION data with replicates, we used DRIMSeq\(^1\). DRIMSeq statistical testing accounted for sequencing and RNA batch according to batch numbers for each sample (Supplementary Table 1). 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 should be from either the CLL SF3B1 WT or SF3B1 K70E condition. A pseudocount of 1 was used for events with 1–2 dropout samples from being excluded from testing. For differential isoform usage testing (FLAIR diffExpr), isoforms were grouped by gene and only genes with at least 25 reads in 4 of the 6 samples were tested. As these libraries were poly(A) selected, we did not distinguish between IR due to incomplete transcript processing and IR deliberately retained due to sample genotype.

**GO analysis.** GO analysis was performed with the R package goseq v1.3.2\(^6\), 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 supplementary Table 1). The expression docount of 1 was used to prevent events with 1 dropout samples from being excluded from testing. For analysis of the PromethION data with replicates, we used DRIMSeq\(^1\). DRIMSeq statistical testing accounted for sequencing and RNA batch according to batch numbers for each sample (Supplementary Table 1). 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 should be from either the CLL SF3B1 WT or SF3B1 K70E condition. A pseudocount of 1 was used for events with 1–2 dropout samples from being excluded from testing. For differential isoform usage testing (FLAIR diffExpr), isoforms were grouped by gene and only genes with at least 25 reads in 4 of the 6 samples were tested. As these libraries were poly(A) selected, we did not distinguish between IR due to incomplete transcript processing and IR 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 ten or more in one of the conditions were plotted. IR were defined as an 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 site\(^6\). 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 v1.3.2\(^6\), 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 supplementary Table 1). The expression docount of 1 was used to prevent events with 1 dropout samples from being excluded from testing. For analysis of the PromethION data with replicates, we used DRIMSeq\(^1\). DRIMSeq statistical testing accounted for sequencing and RNA batch according to batch numbers for each sample (Supplementary Table 1). 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 should be from either the CLL SF3B1 WT or SF3B1 K70E condition. A pseudocount of 1 was used for events with 1–2 dropout samples from being excluded from testing. For differential isoform usage testing (FLAIR diffExpr), isoforms were grouped by gene and only genes with at least 25 reads in 4 of the 6 samples were tested. As these libraries were poly(A) selected, we did not distinguish between IR due to incomplete transcript processing and IR deliberately retained due to sample genotype.
39. Kingan, S. B. et al. A high-quality genome assembly from a single, field-collected spotted lanternfly (Lycorma delicatula) using the PacBio Sequel II System. https://doi.org/10.1101/627679. (2019)
40. Au, K. F., Underwood, J. G., Lee, L. & Wong, W. H. Improving PacBio long read accuracy by short read alignment. 
41. Gordon, S. P. et al. Widespread polycistronic transcripts in fungi revealed by single-molecule mRNA sequencing. PLoS ONE 10, e0132628 (2015).
42. Wan, Y. & Wu, C. J. SF3B1 mutations in chronic lymphocytic leukemia. Blood 121, 4627–4634 (2013).
43. Seki, M. et al. Evaluation and application of RNA-Seq by MinION. DNA Res. 26, 55–65 (2019).
44. Li, H. Minimap and miniasm: fast mapping and de novo assembly for noisy long sequences. Bioinformatics 32, 2103–2110 (2016).
45. Wu, T. D. & Watanabe, C. K. GMAP: a genomic mapping and alignment program for mRNA and EST sequences. Bioinformatics 21, 1859–1875 (2005).
46. Weirather, J. L. et al. Comprehensive comparison of Pacific Biosciences and Oxford Nanopore Technologies and their applications to transcriptome analysis. F1000Res. 6, 100 (2017).
47. Křížanová, K., Echchiki, A., Roux, J. & Šikác, M. Evaluation of tools for long read RNA-seq splice-aware alignment. Bioinformatics 34, 748–754 (2018).
48. Perta, M., Kim, D., Perta, G. M., Leek, J. T. & Salberg, S. L. Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown. Nat. Protoc. 11, 1650–1667 (2016).
49. Yang, C., Chu, J., Warren, R. L. & Biro, I. NanoSim: nanopore sequence read simulator based on statistical characterization. GigaScience 6, 1–6 (2017).
50. Paul, L. et al. SIRVs: Spike-In RNA variants as external isomor controls in RNA-sequencing. https://doi.org/10.1101/080747 (2016).
51. Vaser, R., Soivì, E., Nagarajan, N. & Šikác, M. Fast and accurate de novo genome assembly from long uncorrected reads. Genome Res. 27, 737–746 (2017).
52. Petro, R., Duggal, G., Love, M. I., Irizarry, R. A. & Kingsford, C. Salmon provides fast and bias-aware quantification of transcript expression. Nat. Methods 14, 417–419 (2017).
53. Brooks, A. N. et al. Conservation of an RNA regulatory map between Drosophila and mammals. Genome Res. 21, 193–202 (2011).
54. Nowicka, M. & Robinson, M. D. DRMSeq: a Dírichlet-multinomial framework for multivariate count outcomes in genomics. F1000Res. 5, 1336 (2016).
55. Dvinge, H. & Bradley, R. K. Widespread intron retention diversifies most cancer transcriptomes. Genome Med. 7, 45 (2015).
56. Jung, H. et al. Intron retention is a widespread mechanism of tumor-suppressor inactivation. Nat. Genet. 47, 1242–1248 (2015).
57. Wang, Q. & Rio, D. C. JUM: a computational method for comprehensive annotation-free analysis of alternative pre-mRNA splicing patterns. Proc. Natl Acad. Sci. USA 115, E8181–E8190 (2018).
58. Lupu, C., Zhu, H., Popescu, N. I., Wren, J. D. & Lupu, F. Novel protein ADTRP regulates TFPI expression and function in human endothelial cells in normal conditions and in response to androgen. Blood 118, 4463–4471 (2011).
59. Casadio, A., Hai, J., Marenco, M., Smith, C. W. J. & Eyars, E. Genome-wide association between branch point properties and alternative splicing. PLoS Comput. Biol. 6, e1000166 (2010).
60. Rivas, M. A. et al. Human genomics. Effect of predicted protein-truncating genetic variants on the human transcriptome. Science 348, 666–669 (2015).
61. Lewis, B. P., Green, R. E. & Brenner, S. E. Evidence for the widespread coupling of alternative splicing and nonsense-mediated mRNA decay in humans. Proc. Natl Acad. Sci. USA 100, 189–192 (2003).
62. Sun, S., Zhang, Z., Sinha, R., Karni, R. & Kainer, A. R. SF2/ASF autoregulation involves multiple layers of post-transcriptional and translational control. Nat. Struct. Mol. Biol. 17, 306–312 (2010).
63. Filichkin, S. A. & Mockler, T. C. Unproductive alternative splicing and nonsense mRNAs: a widespread phenomenon among plant circadian clock genes. Biol. Direct 7, 20 (2012).
64. Fu, X.-D. Exploiting the Hidden Treasure of Detained Introns. Cancer Cell 22, 393–395 (2017).
65. Braun, C. J. et al. Coordinated splicing of regulatory detained introns within oncogenic transcripts creates an exploitable vulnerability in malignant glioma. Cancer Cell 22, 411–426.e11 (2017).
66. Ten Hacken, E. et al. Splicing modulation sensitizes chronic lymphocytic leukemia cells to venetoclax by remodeling mitochondrial apoptotic dependencies. JCI Insight 3, e121438 (2018).
67. Schurch, N. J. et al. How many biological replicates are needed in an RNA-seq experiment and which differential expression tool should you use? RNA 22, 839–851 (2016).
68. Nellore, A. et al. Human splicing diversity and the extent of unannotated splice junctions across human RNA-seq samples on the Sequence Read Archive. Genome Biol. 17, 266 (2016).