RESOURCE/METHODOLOGY

Most human introns are recognized via multiple and tissue-specific branchpoints

Jose Mario Bello Pineda[1,2,3,4] and Robert K. Bradley[1,2,3]

[1]Computational Biology Program, Public Health Sciences Division, Fred Hutchinson Cancer Research Center, Seattle, Washington 98109, USA; [2]Basic Sciences Division, Fred Hutchinson Cancer Research Center, Seattle, Washington 98109, USA; [3]Department of Genome Sciences, University of Washington, Seattle, Wasington 98195, USA; [4]Medical Scientist Training Program, University of Washington, Seattle, Wasington 98195, USA

Although branchpoint recognition is an essential component of intron excision during the RNA splicing process, the branchpoint itself is frequently assumed to be a basal, rather than regulatory, sequence feature. However, this assumption has not been systematically tested due to the technical difficulty of identifying branchpoints and quantifying their usage. Here, we analyzed ~1.31 trillion reads from 17,164 RNA sequencing data sets to demonstrate that almost all human introns contain multiple branchpoints. This complexity holds even for constitutive introns, 95% of which contain multiple branchpoints, with an estimated five to six branchpoints per intron. Introns upstream of the highly regulated ultraconserved poison exons of SR genes contain twice as many branchpoints as the genomic average. Approximately three-quarters of constitutive introns exhibit tissue-specific branchpoint usage. In an extreme example, we observed a complete switch in branchpoint usage in the well-studied first intron of *HBB* (β-globin) in normal bone marrow versus metastatic prostate cancer samples. Our results indicate that the recognition of most introns is unexpectedly complex and tissue-specific and suggest that alternative splicing catalysis typifies the majority of introns even in the absence of differences in the mature mRNA.

[*Keywords*: RNA; alternative splicing; branchpoint]

Supplemental material is available for this article.

Received January 22, 2018; revised version accepted March 9, 2018.

RNA splicing proceeds via a two-step process defined by sequential transesterification reactions between three nucleotides: the first nucleotide of the 5′ splice site, the branch nucleotide (branchpoint) upstream of the 3′ splice site, and the last nucleotide of the 3′ splice site. In the first step of splicing, the 2′ OH group of the branchpoint engages in a nucleophilic attack on the phosphate between the upstream exon and the 5′ splice site, forming a 2′–5′ phosphodiester linkage (the "branch") characteristic of the lariat RNA intermediate and releasing the upstream exon. The 3′ OH group of the now-free upstream exon then engages in a nucleophilic attack on the phosphate between the 3′ splice site and the downstream exon, resulting in release of the intronic lariat and exon ligation (for review, see Wahl et al. 2009). The intronic lariat is then linearized via debranching and subsequently degraded.

The branchpoint therefore plays a critical role in RNA splicing catalysis, similar in importance to the splice sites themselves. The branchpoint's biochemical role in the splicing of specific substrate RNAs has been thoroughly studied accordingly. Nonetheless, the identification, selection, and potential regulation of branchpoints remains poorly understood, even relative to other intronic elements such as the polypyrimidine tract or intronic splicing silencers and enhancers that, like the branchpoint, do not appear in the final mRNA product (for review, see Fu and Ares 2014; Scotti and Swanson 2016).

The study of branchpoints has lagged behind the study of other sequence features that define introns and exons for several reasons. Experimentally identifying branchpoints is technically difficult, since lariats exist only as transient low-abundance RNAs. Computationally predicting branchpoints is similarly difficult due to the low information content of the human branchpoint consensus (Zhuang et al. 1989; Kol et al. 2005; Gao et al. 2008; Corvelo et al. 2010). Finally, perhaps because branchpoints are an essential sequence feature required for splicing

Corresponding author: rbradley@fredhutch.org
Article published online ahead of print. Article and publication date are online at http://www.genesdev.org/cgi/doi/10.1101/gad.312058.118. Freely available online through the *Genes & Development* Open Access option.

catalysis (in contrast to other intronic sequence elements that influence splice site recognition but are not universally required for splicing), branchpoints have frequently been assumed to play basal, rather than regulatory, roles.

Nonetheless, several lines of evidence suggest that branchpoint selection may frequently contribute to regulated splice site recognition in human cells. Detailed studies of specific introns revealed that both alternative and constitutive introns may have multiple branchpoints associated with a single 3′ splice site. This branchpoint degeneracy was initially observed in SV40 early pre-mRNA, which is alternatively spliced into the large T and small t mRNAs via an alternative 5′ splice site. This intron contains multiple branchpoints associated with a single 3′ splice site, six of which are used to generate large T mRNAs and one of which is used to generate small t mRNAs (Noble et al. 1987). Multiple branchpoints were subsequently found to be associated with single 3′ splice sites of the adenovirus *E1a* (Gattoni et al. 1988), rat *Tpm1* (Helfman and Ricci 1989), and human *GH1* and *HTR4* (Hartmuth and Barta 1988; Hallegger et al. 2010) genes as well as within a majority of 52 introns of 20 human housekeeping genes (Gao et al. 2008).

Introns with multiple branchpoints can be subject to branchpoint competition during both constitutive and alternative splicing. Studies of a variant of the first intron of β-globin, which was engineered to contain a duplicated branchpoint sequence, revealed that branchpoint competition can occur even within constitutive introns (Zhuang et al. 1989). Branchpoint competition similarly contributes to alternative splicing. The contexts and positions of competing branchpoints can influence the recognition of competing 3′ splice sites (Reed and Maniatis 1988; Smith et al. 1993; Bradley et al. 2012), competing 5′ splice sites (Noble et al. 1988), cassette exons (Kol et al. 2005; Corvelo et al. 2010), and mutually exclusive exons (Mullen et al. 1991; Southby et al. 1999). These previous studies of specific introns suggest that redundant branchpoints may be common, potentially permitting regulated or cell type-specific recognition of many splice sites and introns. However, systematically identifying roles for branchpoint selection in splicing regulation has been hindered by the lack of a genome-wide branchpoint annotation as well as the difficulty in quantifying branchpoint usage.

Recent studies have made significant progress toward generating partial genome-wide branchpoint annotations. Even though lariats are transient RNAs with unusual chemical linkages, they can nonetheless be reverse-transcribed and incorporated into cDNA libraries. The branchpoint associated with a given lariat can then be identified by sequencing the junction between the 5′ splice site and the lariat. Because reverse transcriptase frequently incorporates a mismatch, insertion, or deletion when traversing the 2′–5′ phosphodiester linkage at the branch, the precise branchpoint location can be mapped by identifying putative 5′ splice site–branchpoint junctions where the sequenced cDNA has a mismatch specifically at the inferred branchpoint location (Vogel et al. 1997; Gao et al. 2008). Taggart et al. (2012) exploited the occasional incorporation of lariats into cDNA libraries to perform the first de novo branchpoint identification using RNA sequencing (RNA-seq), identifying 862 branchpoints in the human genome. Mercer et al. (2015) later created lariat-enriched cDNA libraries with RNase R digestion and targeted RNA recovery to identify 59,359 branchpoints, the largest annotation to date. These genome-wide studies identified a minority of introns with multiple branchpoints (9% in Taggart et al. [2012] and 32% in Mercer et al. [2015]). However, most of those branchpoints were annotated based on just one or a few sequenced lariats. Therefore, it is possible that undersampling of lariats resulted in underestimates of branchpoint multiplicity in those previous studies. Furthermore, because only one or a few lariats were observed for most branchpoints (an inevitable consequence of the very low abundance of lariat RNA species), quantitative estimates of branchpoint usage remain elusive.

Here, we sought to determine whether branchpoints are typically just basal sequence features of introns or whether branchpoint recognition is frequently complex or regulated. We performed a very large-scale analysis to systematically identify branchpoints and quantify their usage across diverse human tissues in both normal and diseased states. Our results indicate that almost all human introns have multiple branchpoints, which are frequently used in a tissue-specific manner. Branchpoint abundance correlates with alternative splicing. Our data demonstrate that branchpoint recognition is unexpectedly complex, giving rise to cell type-specific splice site recognition during both constitutive and alternative splicing.

Results

A large-scale analysis of RNA-seq data enables global branchpoint annotation

We sought to create a genome-wide branchpoint annotation by taking advantage of the occasional reverse transcription of lariats and their subsequent incorporation into cDNA libraries. The transient nature of lariat RNA as well as the specific selection of polyadenylated RNA in many RNA-seq library construction protocols render lariat incorporation rare. Informative reads from lariats—those that span the junction between the 5′ splice site and a branchpoint rather than simply lying within the intron—are even rarer. To address this statistical challenge, we performed an extremely large-scale analysis of ∼1.31 trillion reads from 17,164 RNA-seq data sets (Supplemental Table S1). These data sets were generated from healthy as well as diseased tissues, including ∼550 billion reads from the Genotype-Tissue Expression (GTEx) project's survey of healthy tissues (Melé et al. 2015) and ∼490 billion reads from The Cancer Genome Atlas's survey of primary and peritumoral tissues from diverse cancers. Together, these 17,164 data sets represent a comprehensive survey of cell types and physiological states.

For each RNA-seq data set, we identified lariat-derived reads that spanned 5′ splice site–branchpoint junctions by sequentially aligning reads to the transcriptome,

genome, 5′ splice sites, and 3′ splice sites (Fig. 1A; Supplemental Fig. S1). This alignment strategy was modeled after the "split-read" approach used by Mercer et al. (2015). In brief, we first prefiltered each RNA-seq data set by removing all reads that aligned to the transcriptome or genome. We then aligned the remaining reads to a database of all annotated 5′ splice sites, requiring a minimum of 20 nucleotides (nt) of aligned sequence. We trimmed each read alignment to remove the 5′ splice site sequence and then aligned each trimmed read to a database of all annotated 3′ splice sites, requiring complete alignment of the trimmed read with a minimum of 20 nt of aligned sequence within 250 nt of the 3′ splice site itself. We then restricted to reads that aligned to 5′ and 3′ splice sites within a single gene in the "inverted" pattern (e.g., where the end of the read maps upstream of the start of the read) expected of reads arising from lariats rather than linear introns. The inferred branchpoint location is then the last nucleotide of the alignment of the trimmed read to the 3′ splice site. Finally, in order to obtain nucleotide-level resolution of branchpoint locations, we restricted to reads with a mismatch at the inferred branchpoint location. Such mismatches are strongly associated with correctly inferred branchpoints, as reverse transcriptase frequently incorporates an incorrect nucleotide when traversing the 2′–5′ phosphodiester linkage at the branch (Vogel et al. 1997; Gao et al. 2008).

Manual inspection of the resulting data set revealed that we comprehensively annotated 5′ splice site–branchpoint pairs for many genes. For example, we identified branchpoints within all but one intron of the gene encoding the splicing factor SRSF5. Our 5′ splice site–branchpoint pairs revealed complex splicing patterns for *SRSF5*, including alternative 5′ splice site usage, skipping of multiple cassette as well as constitutive coding exons, and usage of branchpoints that were proximal as well as distal to 3′ splice sites (Fig. 1B).

Comparison with published branchpoint annotations

As the split-read alignment procedure can be confounded by gene duplications or the presence of other repetitive genomic DNA, we assigned a confidence level to each inferred 5′ splice site–branchpoint pair. For a 5′ splice site–

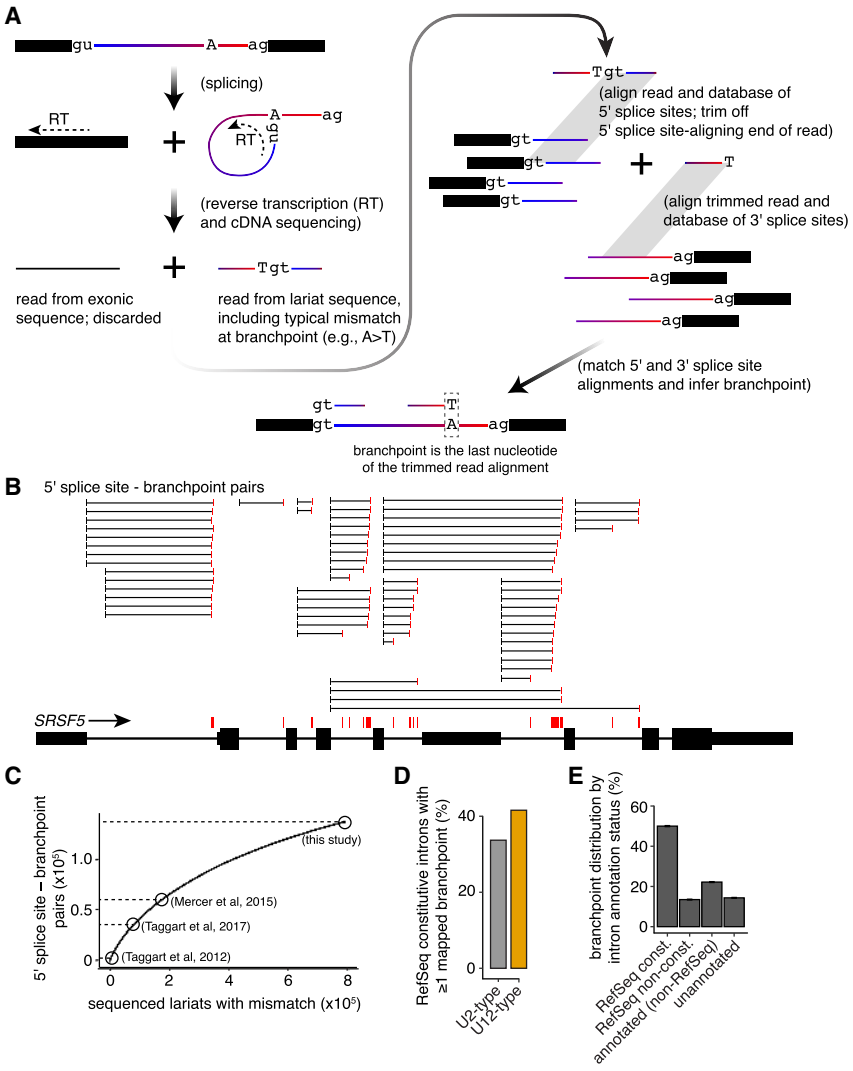

Figure 1. Genome-wide branchpoint annotation from RNA-seq data. (*A*) Overview of our branchpoint detection algorithm (see also Supplemental Figure S1). (*B*) Branchpoint annotation of *SRSF5*. For simplicity, only the intron-distal splice site of a competing 5′ splice site event within the first intron is illustrated in the exon–intron structure. (Vertical red bars) Branchpoints; (horizontal black lines) 5′ splice site–branchpoint pairs. The plot is based on an image from the University of California at Santa Cruz (UCSC) Genome Browser (Meyer et al. 2013). (*C*) Branchpoint detection rate as a function of the number of sequenced lariats. We randomly sampled from all sequenced lariats analyzed in our study and computed the number of distinct 5′ splice site–branchpoint pairs detected. As 5′ splice site–branchpoint pairs were not reported by other studies, we illustrated the number of reported branchpoints instead. For Taggart et al. (2017), we illustrated their "high-confidence" set of branchpoints. (*D*) Fraction of all RefSeq constitutive introns with one or more mapped branchpoints. (*E*) Distribution of mapped branchpoints among different annotation classes. (RefSeq const.) RefSeq constitutive introns; (RefSeq non-const.) RefSeq nonconstitutive introns; [annotated (non-RefSeq)] introns present in the UCSC, Ensembl, or Mixture of Isoforms (MISO) annotation databases but not RefSeq; (unannotated) introns formed by unannotated ligation of annotated 5′ and 3′ splice sites.

branchpoint pair to meet the highest confidence level, we required (1) that ≥5% of supporting reads have mismatches at the branchpoint but no other mismatches or indels (insertions/deletions) in the 3′ splice site alignment and (2) that the 25 nt downstream from the 5′ splice site, 25 nt upstream of the inferred branchpoint, and 25 nt of the lariat centered on the inferred branchpoint all be unique (not present in the transcriptome or genome). We successively relaxed these criteria for the moderate and low confidence levels. We removed the sequence uniqueness criteria for moderate-confidence branchpoints and allowed additional mismatches and indels in the 3′ splice site alignment for low-confidence branchpoints. We identified a total of 136,998, 9182, and 48,935 5′ splice site–branchpoint pairs at high, moderate, and low confidence levels. Our branchpoint annotations were robust with respect to the specific details of the thresholds used. For example, requiring that ≥25% of supporting reads have mismatches at the branchpoint but not other mismatches or indels in the 3′ splice site alignment (a fivefold increase in stringency) resulted in only a 2.9% decrease in the number of high-confidence 5′ splice site–branchpoint pairs that we identified.

We next assessed the likely accuracy of our branchpoint inference procedure for each confidence level. Biochemical studies and lariat sequencing have revealed that adenosine is the most effective and frequent branchpoint ribonucleotide (Gao et al. 2008), suggesting that global branchpoint adenine frequency correlates with inference accuracy. Branchpoints that we identified at high, moderate, and low confidence levels had adenine frequencies of ∼77%, 50%, and 32%, indicating that our confidence levels correlate with likely inference accuracy. Therefore, we restricted all subsequent global analyses to 5′ splice site–branchpoint pairs detected at the highest confidence level.

We next compared our branchpoint annotations with previously published branchpoint data sets (Table 1). We identified 70,935 and 94,216 more branchpoints than were reported in Mercer et al. (2015) and Taggart et al. (2017), the largest sets of branchpoint annotations published to date. (For comparison with Taggart et al. [2017], we used their "high-confidence" set of branchpoints.) Our annotation exhibited a branchpoint adenine frequency of ∼77% versus 78% and 55% for Mercer et al. (2015) and Taggart et al. (2017). The lower adenine frequency for the annotation of Taggart et al. (2017) may be due to differences in the methods that each study used to call branchpoints. Like Mercer et al. (2015), we restricted to reads with a mismatch at the inferred branchpoint, which is diagnostic of reverse transcriptase incorporating an incorrect nucleotide when traversing the 2′–5′ phosphodiester linkage at the branch (Vogel et al. 1997; Gao et al. 2008). In contrast, Taggart et al. (2017) did not require a mismatch at the inferred branchpoint. Instead, they aligned putative branchpoint sequence contexts to the U2 small nuclear RNA (snRNA) sequence and called branchpoints at the inferred bulged nucleotide (Taggart et al. 2017).

Parent gene expression and intron length determine branchpoint detection rate

Despite the extremely large-scale nature of our analysis, we did not approach saturation. We estimated the branchpoint detection rate as a function of the number of sequenced lariats by randomly sampling from all branchpoint-spanning reads. Even though we detected many more branchpoints than did previous studies, branchpoint detection continued to increase rapidly as a function of the number of sequenced lariats throughout the dynamic range of our study (Fig. 1C).

We detected one or more branchpoints within ∼37% and 42% of U2- and U12-type constitutive introns present in the RefSeq annotation (O'Leary et al. 2016), where we defined constitutive introns as those that were present in all child transcripts of a given RefSeq gene (Fig. 1D). Fifty percent of detected branchpoints fell within RefSeq constitutive introns, while 35% fell within nonconstitutive introns present in the RefSeq, University of California at Santa Cruz (UCSC), Ensembl, or Mixture of Isoforms (MISO) isoform databases (Fig. 1E; Katz et al. 2010; Flicek et al. 2013; Meyer et al. 2013). An unexpectedly large percentage (14%) of 5′ splice site–branchpoint pairs corresponded to introns that were not annotated in any of those isoform databases, resulting from skipping of one or more ostensibly constitutive exons. While some such cases may correspond to stable isoforms with potential cellular functions, many may simply represent by-products of splicing mistakes.

Table 1. *Comparison of published branchpoint annotations*

Study	RNA-seq reads analyzed	Branchpoints	5′ splice site–branchpoint pairs	A frequency at branchpoint
Gao et al. 2008	NA	60	60	85%
Taggart et al. 2012	∼1.2 billion	862	Not reported	39%
Mercer et al. 2015	∼3 billion	59,359	Not reported	78%
Taggart et al. 2017	∼11.3 billion	36,078	Not reported	55%
This study: high confidence	∼1.31 trillion	130,294	136,998	77%
This study: moderate confidence	∼1.31 trillion	8220	9182	50%
This study: low confidence	∼1.31 trillion	47,894	48,935	32%

The high-, moderate-, and low-confidence categories used in our study are mutually exclusive. The numbers for Taggart et al. (2017) correspond to their "high-confidence" set of branchpoints.

The numbers of branchpoints that we detected per intron or gene were highly variable. We obtained seemingly near-complete annotations for some genes (e.g., *SRSF5* in Fig. 1B) and few or no branchpoints for any introns of other genes. This high level of variability in branchpoint detection could arise from many factors, including differences in parent gene expression, intron length, and lariat stability. We tested whether each of these factors contributed to differences in branchpoint detection rate. We restricted these power analyses to constitutive introns in order to avoid additional complexities arising from alternative splicing. The branchpoint detection rate was strongly positively and negatively correlated with parent gene expression and intron length, as expected from random sampling of lariats (Supplemental Fig. S2A–C). Branchpoints from very short (<200-base-pair [bp]) introns were underrepresented, presumably because most RNA-seq protocols intentionally deplete such short RNAs during library preparation.

We tested whether some lariats had unusually short or long half-lives by comparing the observed abundance of each lariat with its expected abundance, defined as the ratio of its parent gene expression to intron length. The distribution of observed to expected abundances followed a normal distribution across all sequenced lariats, consistent with a model in which lariats are degraded randomly. Lariats with a guanine branchpoint exhibited a 1.6-fold greater abundance than expected, suggesting that they are frequently more stable than lariats with adenine, cytosine, or thymine/uracil branchpoints (Supplemental Fig. S2D). These findings are consistent with a previous report that lariats formed from nonadenine mutants of the rabbit *HBB* gene were resistant to debranching relative to lariats formed via the wild-type adenine branchpoint (Hornig et al. 1986).

Distal branchpoints contribute to alternative exon and intron recognition

We used our genome-wide branchpoint annotation to identify sequence features contributing to branchpoint recognition and usage. Branchpoints within constitutive introns were most frequently adenine (82.5%), followed by guanine (8.7%), thymine/uracil (4.7%), and cytosine (4.1%). We observed a modest preference for thymine/uracil at the −2 position relative to the branchpoint, as reported previously (Gao et al. 2008). However, this preference was restricted to adenine branchpoints, with no site-specific sequence preferences at any other nucleotides for nonadenine branchpoints (Fig. 2A).

Branchpoints exhibited a tightly constrained spatial distribution, as reported by previous studies (Taggart et al. 2012, 2017; Mercer et al. 2015). Branchpoints within U2-type constitutive introns were positioned at a median of 28 nt upstream of the 3′ splice site, with 80% of such branchpoints found within the positions −49 and −20 nt. Branchpoints within U12-type introns exhibited a bimodal distribution (Fig. 2B). Approximately half of such U12-type branchpoints were found in close proximity (within 20 nt) of the 3′ splice site, as observed previously (Dietrich et al. 2001; Taggart et al. 2017). In contrast, approximately half of U12-type introns were located only modestly closer to the 3′ splice site than we observed for U2-type branchpoints. We classified introns as U2- or U12-type by finding the best match between each 5′ splice site sequence to the U2- and U12-type consensus sequences. The U2- and U12-type 5′ splice site consensus sequences are distinct (Sheth et al. 2006), making frequent misclassification unlikely. However, we cannot rule out the possibility that classification error contributes to the unexpected bimodal spatial distribution for U12-type branchpoint positions.

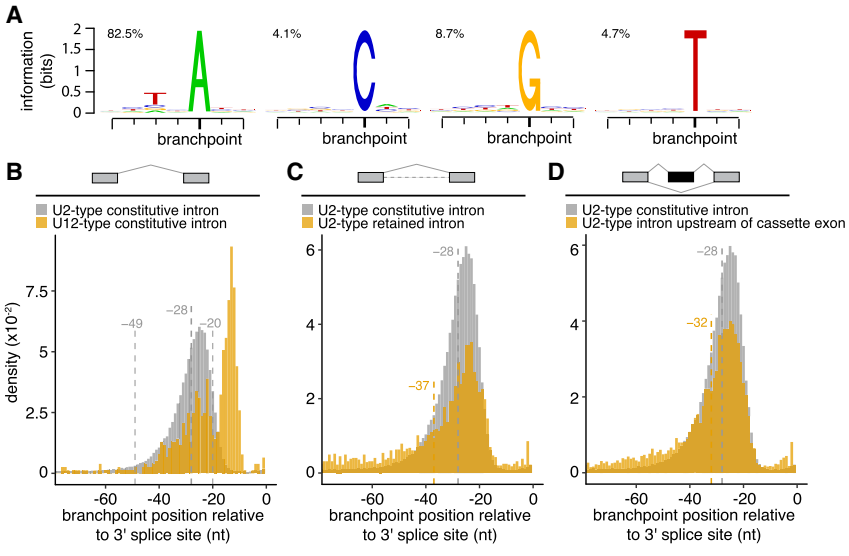

Figure 2. Branchpoint position, but not sequence context, is constrained. (*A*) Sequence logos of branchpoint contexts. The plot is restricted to branchpoints within RefSeq constitutive introns. (*B*) Histogram of branchpoint positions relative to the 3′ splice site, where position −1 nt corresponds to the last intronic nucleotide. Vertical dashed lines at −20, −28, and −49 nt illustrate the 10th, 50th, and 90th percentiles of positions for U2-type introns. The plot is restricted to branchpoints within RefSeq constitutive introns. (*C*) As in *B* but for U2-type introns classified as constitutive or retained. To ensure that the analyzed sets of introns were disjoint, we restricted to constitutive introns that did not overlap introns annotated as potentially retained in the MISO version 2.0 annotation even if those introns did not exhibit retention in our data. The vertical dashed line at −28 nt illustrates the median position for constitutive introns. (*D*) As in *B* but for U2-type introns classified as constitutive or upstream of a cassette exon. To ensure that the analyzed sets of introns were disjoint, we restricted to constitutive introns that did not overlap introns associated with cassette exons even if those cassette exons did not exhibit alternative splicing in our data. The vertical dashed line at −28 nt illustrates the median position for constitutive introns.

While most branchpoints were positioned proximal to the 3′ splice site, a subset was located further upstream. Distal branchpoints, located ≥50 nt upstream, constituted only 9.5% of branchpoints in U2-type constitutive introns (Fig. 2B). In contrast, distal branchpoints frequently occurred in introns associated with alternative splicing events, consistent with previous reports (Corvelo et al. 2010; Taggart et al. 2012, 2017). We quantified alternative splicing across 16 human tissues and restricted to "switch-like" events that exhibited changes in isoform ratio ("switch scores") of ≥25% between tissues. This restriction focused our analysis on regulated tissue-specific splicing rather than low-abundance isoforms that might result from stochastic splicing. Distal branchpoints occurred at frequencies of 39.5% and 28.7% within U2-type introns that were frequently retained or positioned upstream of cassette exons (Fig. 2C,D). Far-distal branchpoints, located ≥100 nt upstream, occurred at frequencies of 4.6%, 22.0%, and 13.9% in U2-type constitutive introns, retained introns, and introns upstream of cassette exons. This unexpectedly strong enrichment for distal and far-distal branchpoints in switch-like retained introns strongly suggests that branchpoint position contributes to regulated intron recognition.

Almost all constitutive introns have multiple branchpoints

We anecdotally noticed that many introns contained multiple annotated branchpoints. This branchpoint multiplicity was common even in constitutive introns, which are not subject to alternative splice site usage yet frequently contain an unexpectedly large number of branchpoints. Given this surprising degree of branchpoint multiplicity, we sought to confirm the results of our high-throughput branchpoint inference procedure with direct lariat sequencing. We selected four constitutive introns within *MBNL1*, *POLR3A*, *SNX9*, and *VASP*, each of which exhibited high branchpoint multiplicity, with six or seven branchpoints discovered within RNA-seq libraries from the K562 erythroleukemic cell line alone. We generated cDNA libraries from K562 cell lysate, used nested PCR to specifically amplify lariats from each of those four introns, performed Sanger sequencing on single amplicons with colony sequencing, and inferred branchpoints from each sequenced amplicon (Supplemental Fig. S3A,B). For each intron, we validated the majority of computationally inferred branchpoints and furthermore discovered new branchpoints (Fig. 3A,B). In addition to experimentally confirming the striking branchpoint multiplicity that we inferred for many introns, these results demonstrated that many or most introns are still undersampled despite our very large-scale RNA-seq analysis.

We next tested whether branchpoint multiplicity was an unusual feature of specific introns or was instead a common characteristic of many introns. Accurately estimating branchpoint abundance is challenging for two reasons. First, as revealed by our power analysis and targeted lariat sequencing experiments, our study has not approached saturation of lariat sequencing even for introns with many annotated branchpoints (Figs. 1C, 3A,B). Second, our lariat sequencing depth varied by orders of magnitude for different introns.

We simultaneously controlled for both of those effects by stratifying all analyses by per-intron lariat sequencing depth. Simply binning each intron according to the number of sequenced lariats revealed that the vast majority of constitutive introns contained multiple lariats. Ninety-five percent of constitutive introns with the greatest sequencing depth (≥250 sequenced lariats) contained two or more distinct branchpoints, with a mean of 6.75 branchpoints detected per intron (Fig. 3C,D). Sequencing just five to 10 lariats per intron was sufficient to detect multiple branchpoints in the majority of introns. Ninety-five percent is probably an accurate estimate of the fraction of constitutive introns with multiple branchpoints, as an asymptote is clearly evident in our power analysis (Fig. 3C). In contrast, additional lariat sequencing will probably reveal novel branchpoints for the 95% of introns exhibiting branchpoint multiplicity (no asymptote is visible in the relevant power analysis) (Fig. 3D), consistent with our discovery of novel branchpoints via direct lariat sequencing of *MBNL1*, *POLR3A*, *SNX9*, and *VASP* introns.

Our estimates of branchpoint multiplicity could potentially be confounded by small nontemplated insertions or deletions generated by reverse transcriptase when traversing the 2′–5′ phosphodiester linkage at the branch (Vogel et al. 1997; Gao et al. 2008). Deletions do not confound our analysis, as they do not result in a mismatch at the branchpoint. However, insertion of a single nucleotide could result in incorrect inference of a branchpoint at the +1 position with respect to the actual branchpoint. (Insertion of two or more nucleotides, which is relatively infrequent, would result in multiple mismatches. Such reads would not satisfy our criteria for high-confidence branchpoints, except for the unlikely case where the randomly inserted nucleotides matched the genomic sequence.) To test whether our branchpoint multiplicity estimates were biased by this potential source of error, we took the conservative approach of collapsing all adjacent branchpoints into a single branchpoint. Even after applying this merge procedure, we estimated that ~94% of constitutive introns contained multiple branchpoints, with a mean of five branchpoints per intron for introns with the most lariat sequencing coverage (Supplemental Fig. S4A,B). We conclude that high branchpoint multiplicity typifies the vast majority of human introns.

Branchpoint position strongly influences branchpoint usage

We next attempted to identify sequence features that contribute to basal branchpoint recognition and selection in the face of high branchpoint multiplicity. We took advantage of the large-scale nature of our study to quantitatively estimate branchpoint usage across 54 healthy human tissues. We removed transcriptome- or genome-aligning reads from the ~550 billion reads sequenced by the GTEx project (Melé et al. 2015) and aligned the remaining reads to all lariat sequences (5′ splice site–branchpoint

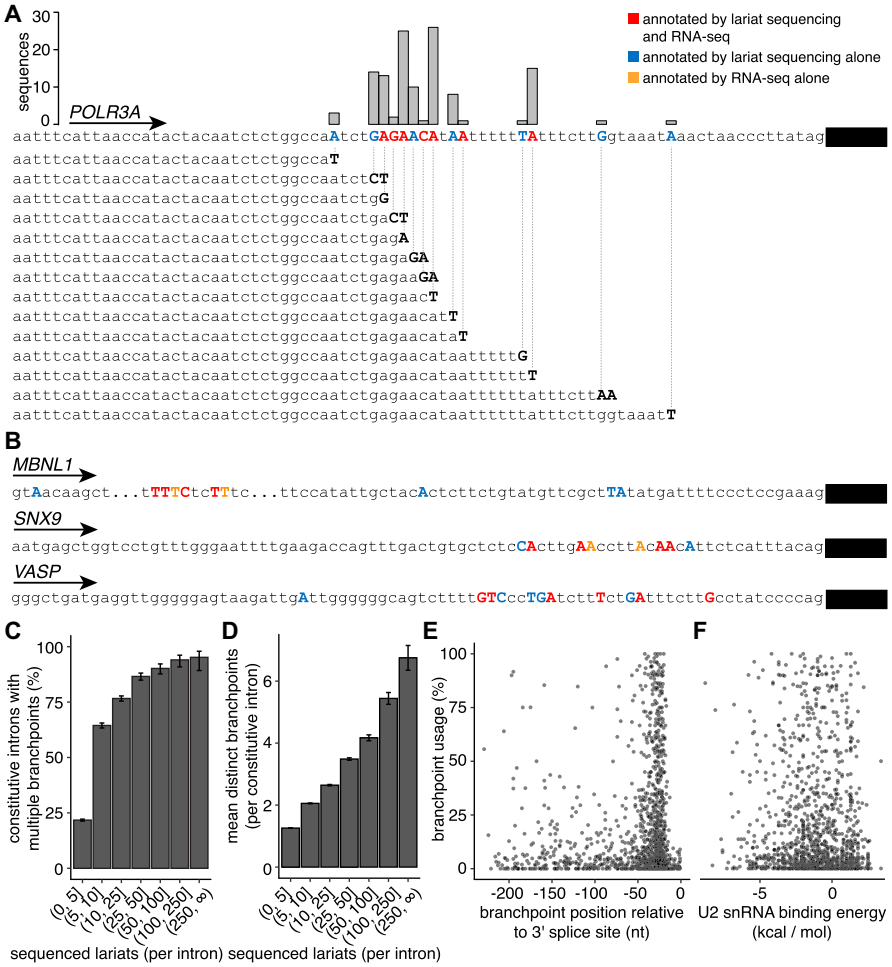

Figure 3. Most constitutively spliced introns contain multiple branchpoints. (*A*,*B*) Branchpoint annotations of introns within *POLR3A* (*A*) and *MBNL1*, *SNX9*, and *VASP* (*B*) based on RNA-seq analysis as well as direct lariat sequencing. Colors indicate the evidence supporting each branchpoint. Examples of sequenced lariats are shown for *POLR3A*. (*C*) The fraction of constitutive introns with multiple branchpoints as a function of the number of sequenced lariats with a mismatch at the branchpoint. Error bars indicate 95% confidence interval estimated with a proportion test. (*D*) As in *C* but illustrating the mean number of branchpoints per intron. Error bars indicate standard deviation of the mean, estimated by bootstrapping. (*E*) Branchpoint usage as a function of the relative branchpoint position. Branchpoint usage is defined as the number of sequenced lariats supporting a given 5′ splice site–branchpoint pair divided by the total number of sequenced lariats mapped to that 5′ splice site. Each point corresponds to a single branchpoint. The plot is restricted to constitutive introns with two or more branchpoints. The two most commonly used branchpoints per intron are illustrated. (*F*) As in *E* but illustrating estimated binding energy to the U2 snRNA sequence AUGAUGUG for each branchpoint context.

pairs). We restricted to reads with a lesion (mismatch or small indel) specifically at the branchpoint in order to help ensure that the reads originated from reverse transcription of branched RNA. For each tissue, we collated reads sampled from different individuals in order to increase our lariat sequencing depth. We then estimated branchpoint usage by computing the frequency with which a particular 5′ splice site–branchpoint pair was used relative to all branchpoints associated with that 5′ splice site. As we sought to identify sequence features that influenced basal branchpoint recognition independent of potential *cis*- or *trans*acting regulation, we estimated basal branchpoint usage by averaging branchpoint usage across all 54 tissues.

We focused on the two key features that define a branchpoint: its location relative to the 3′ splice site and its complementarity to the U2 snRNA sequence. We restricted to U2-type introns and focused our analysis on the two most frequently used branchpoints within each intron. Plotting quantitative branchpoint usage as a function of branchpoint position revealed that the majority of most frequently used branchpoints resided within a narrow window, consistent with the restricted genome-wide distribution of all branchpoint positions (Fig. 3E). While a few far-distal branchpoints were predominantly used, such examples were relatively uncommon.

In contrast to branchpoint position, complementarity to the U2 snRNA was not strongly correlated with

branchpoint usage. We computed the binding energy of each branchpoint sequence context to the U2 snRNA sequence AUGAUGUG, with the exception of the branchpoint itself, which appears as a bulge in the structure. While previous studies have shown that U2 snRNA complementarity is associated with branchpoint recognition, this association is very weak (Mercer et al. 2015). Consistent with previous results, we observed little association between U2 snRNA complementarity and quantitative branchpoint usage (Fig. 3F). Many branchpoints were very poor matches to the U2 snRNA yet were predominantly used. In contrast, most branchpoints within U12-type constitutive introns were comparatively better matches to the U12 snRNA sequence AGGAAUG (Supplemental Fig. S4C).

Highly regulated ultraconserved introns have an unusually high number of branchpoints

Since constitutive introns exhibited such a surprising degree of branchpoint multiplicity, we hypothesized that introns that were associated with alternative splicing might exhibit even more. Manual inspection of specific introns flanking highly regulated cassette exons, such as the "poison" exons of *SRSF5* and *SRSF3*, supported this hypothesis (Figs. 1B, 4A). *SRSF5* and *SRSF3* are members of the SR gene family, each of which contains a highly regulated "poison" splicing event that introduces an in-frame premature termination codon into the mature transcript. Poison exons contribute to SR splicing factor homeostasis and overlap with ultraconserved or highly conserved genomic sequence (Lareau et al. 2007; Ni et al. 2007).

We first confirmed that the large branchpoint cluster upstream of the *SRSF3* poison exon was correctly annotated with direct lariat sequencing. As with our studies of constitutive introns, direct lariat sequencing of the *SRSF3* intron both confirmed the computationally inferred branchpoint cluster and revealed novel branchpoints (Fig. 4B). These branchpoints were spread throughout the highly conserved intronic region upstream of the poison exon, suggesting that they likely contribute to the purifying selection acting on this genomic sequence.

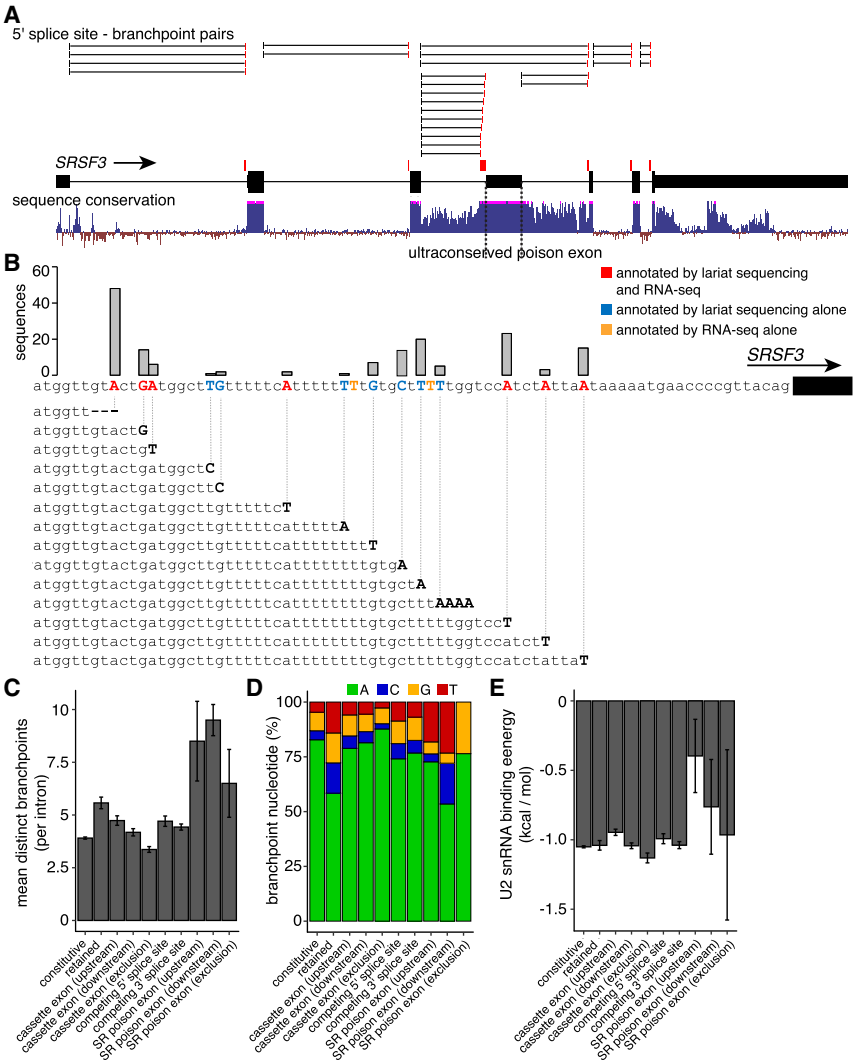

Figure 4. Regulated alternative splicing is associated with high branchpoint multiplicity. (*A*) Branchpoint annotation for *SRSF3*. Sequence conservation was performed with phastCons 100-vertebrate conservation track (Siepel et al. 2005). The plot was based on an image from the UCSC Genome Browser (Meyer et al. 2013). (*B*) Branchpoint annotation for the intron upstream of the *SRSF3* poison exon, based on RNA-seq analysis as well as direct lariat sequencing. Colors indicate the evidence supporting each branchpoint. (*C*) The mean number of branchpoints detected in each of the illustrated classes of introns. Alternative splicing annotations were based on the MISO version 2.0 isoform database (Katz et al. 2010). The plot is restricted to introns with ≥25 sequenced lariats to help control for intron-specific variability in lariat sequencing depth. Error bars indicate standard deviation of the mean, estimated by bootstrapping. (*D*) As in *C* but illustrating the frequencies with which each branchpoint nucleotide occurs. (*E*) As in *C* but illustrating the mean estimated U2 snRNA-binding energy. Error bars indicate standard deviation of the mean, estimated by bootstrapping.

We next tested whether branchpoint abundance was associated with alternatively spliced sequences at a genome-wide level. We classified introns as constitutive, retained, upstream of or downstream from cassette exons, containing a cassette exon, or containing competing 5′ or 3′ splice sites. We considered introns that were associated with poison exons of SR genes as a distinct class. All introns associated with the inclusion of alternatively spliced sequence were enriched for branchpoints relative to constitutive introns, with retained introns displaying the greatest enrichment (~43%) (Fig. 4C). While introns upstream of as well as downstream from cassette exons were enriched for branchpoints, introns corresponding to cassette exon exclusion exhibited a modest depletion relative to constitutive introns, suggesting that some branchpoints downstream from cassette exons are used only in the context of exon inclusion. (This comparison was made possible by our enumeration of 5′ splice site–branchpoint pairs rather than branchpoints alone.) We observed the same trend, although with much greater branchpoint multiplicity, for introns associated with poison exons of SR genes.

In addition to exaggerated multiplicity, branchpoints within introns that were associated with alternative splicing exhibited other unusual characteristics. Adenine is found at ~83% of branchpoints within constitutive introns but only ~58% of branchpoints within retained introns (Fig. 4D). Adenine frequencies are highest for 5′ splice site–branchpoint pairs corresponding to cassette exon exclusion (~88%), mirroring our observation that cassette exon exclusion is associated with reduced branchpoint multiplicity even relative to constitutive intron splicing. Introns flanking the poison exons of SR genes contained branchpoints within sequence contexts that were unusually poor matches to the U2 snRNA consensus, with an average of −0.4 kcal/mol for introns upstream of SR poison exons versus −1.1 kcal/mol for constitutive introns (Fig. 4E). Together, our data indicate that an abundance of branchpoints, many of which are suboptimal, likely contributes to regulated alternative splicing.

Branchpoint usage is frequently tissue-specific

The branchpoint multiplicity that characterizes most introns theoretically permits tissue-specific branchpoint selection and intron recognition even for constitutive introns. We anecdotally noticed a striking example of this within the first intron of *HBB* (encoding β-globin), a well-studied splicing substrate. Early biochemical studies demonstrated that *HBB*'s first intron forms a lariat RNA via an adenine branchpoint at position −37 nt relative to the 3′ splice site (Ruskin et al. 1984). Mutating this branchpoint to a guanine did not abolish in vitro splicing of its parent intron. Instead, branchpoint usage shifted to a 3′ splice site-proximal adenine located at position −24 nt (Ruskin et al. 1985). While the dominant branchpoint at position −37 nt is an excellent match to the U2 snRNA, with a binding energy of −5.2 kcal/mcol, the cryptic branchpoint at position −24 nt has a binding energy of just −0.5 kcal/mol. This difference may explain why the downstream branchpoint was used in vitro only when the dominant branchpoint was mutated.

Given these biochemical studies, we expected to observe exclusive usage of the −37-nt branchpoint in our own data for *HBB*. Unexpectedly, we found that branchpoint usage was instead highly tissue-specific, with nonoverlapping sets of branchpoints used in blood versus metastatic prostate cancer (Fig. 5A). The −37-nt branchpoint was present in 79% of lariats sequenced from normal or leukemic peripheral blood or bone marrow, with infrequent usage of other branchpoints at positions −78, −41, and −24 nt. In contrast, in metastatic prostate cancer samples, branchpoints at positions −30 nt and −26 nt constituted 31% and 58% of branchpoint usage. The −37-nt branchpoint was virtually unused.

Since we observed such striking variation in branchpoint usage even within the well-studied first intron of *HBB*, we hypothesized that tissue-specific branchpoint usage might be more common than is currently recognized. We therefore sought to use direct lariat sequencing to identify differentially used branchpoints within the *VASP* and *SRSF3* introns studied above as exemplars of constitutive and alternative splicing. We first confirmed that our direct lariat sequencing protocol was sufficiently reproducible to quantify differential branchpoint usage. We amplified, cloned, and sequenced lariats from the *VASP* and *SRSF3* introns in technical duplicates from five tissues obtained from healthy donors (peripheral blood, cerebellum, spleen, fetal spleen, and testis). We estimated false discovery rates (FDRs) for each intron by measuring the frequencies of differential branchpoint usage between the technical replicates for each of the [(number of branchpoints) × (number of tissues)] trials, where we defined differential branchpoint usage as differences in usage of ≥10% with a *P*-value of ≤0.01. We estimated FDRs of 2.5% and 2.4% for *VASP* and *SRSF3*, indicating that our assay is robust with respect to experimental variability.

We therefore used our lariat sequencing assay to quantify branchpoint usage for all branchpoints within the *VASP* and *SRSF3* introns. We observed frequent differential branchpoint usage for both introns, including differences between tissues as well as between fetal and adult samples from the same tissue (Fig. 5B,C). Both the *VASP* and *SRSF3* introns contained "switch-like" branchpoints, which were never used in some tissues but were used frequently in others. The four and five differentially used branchpoints that we detected in *VASP* and *SRSF3* far exceeded the number expected from experimental variability alone, with associated *P*-values of 5.6×10^{-3} and 1.9×10^{-5} (Fig. 5D).

We next extended our targeted experimental analysis of *VASP* and *SRSF3* to a genome-wide RNA-seq-based measurement of differential branchpoint usage. We estimated branchpoint usage across healthy human tissues in the GTEx data set and performed a power analysis similar to our approach for estimating genome-wide branchpoint multiplicity (Fig. 3C,D). We focused our analysis on constitutive introns, since branchpoint multiplicity and differential branchpoint usage in the context of constitutive splicing was so unexpected. We restricted to constitutive

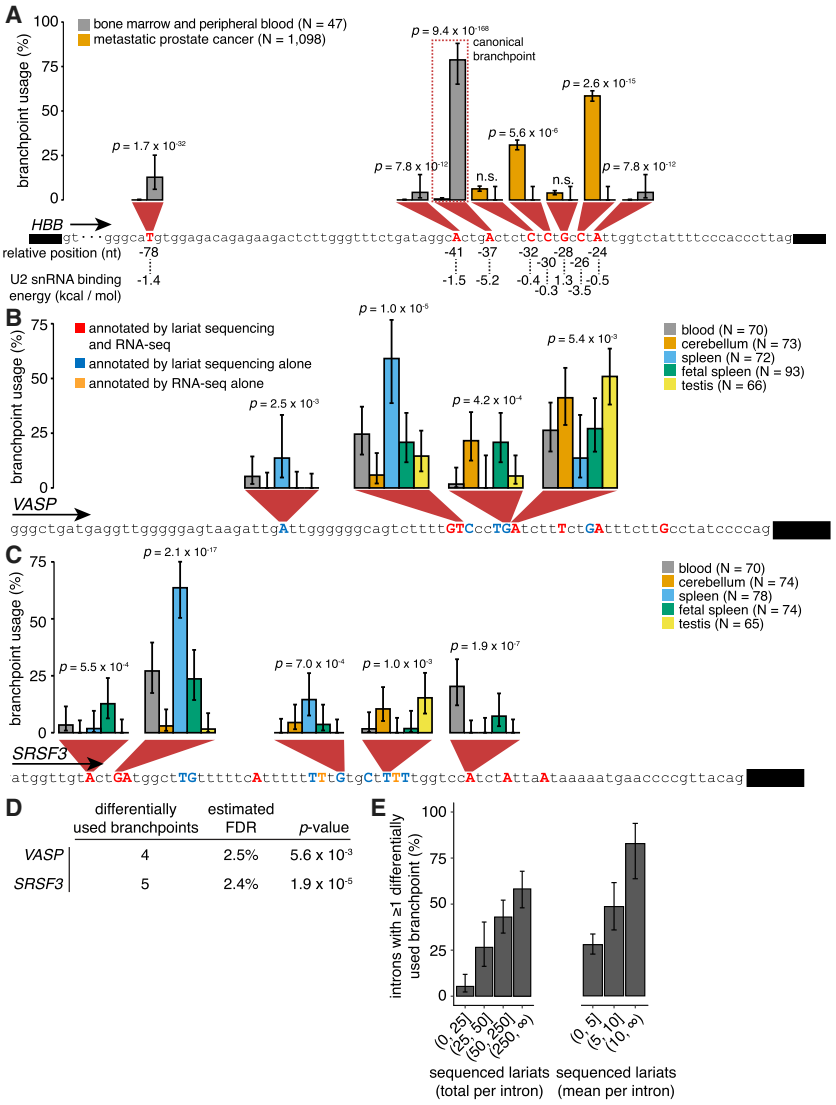

Figure 5. Tissue-specific branchpoint usage is common. (*A*) Branchpoint annotation and estimated branchpoint usage for the first intron of *HBB*. (N) Number of sequenced lariats with a mismatch at the inferred branchpoint. Error bars indicate 95% confidence intervals estimated with the binomial proportion test. *P*-values were estimated with the binomial proportion test. Branchpoints at positions −32 nt, −37 nt (the canonical branchpoint annotated biochemically) (Ruskin et al. 1984), and −41 nt were annotated with moderate, rather than high, confidence due to the nonuniqueness of the *HBB* intronic sequence. (*B*,*C*) As in *A* but for the indicated introns of *VASP* (*B*) and *SRSF3* (*C*). Data are from direct lariat sequencing. *P*-values were estimated with the multinomial proportion test. The plot is restricted to branchpoints exhibiting differential branchpoint usage across the indicated samples, defined as a tissue-specific difference in branchpoint usage of ≥10% with an associated *P*-value ≤0.01 (two-sided test for difference in proportion). The illustrated percentages do not add up to 100% because the plot is restricted to differentially used branchpoints. (*D*) Detection of tissue-specific branchpoint usage in *VASP* and *SRSF3* relative to the empirical false discovery rate (FDR) for each intron. Empirical FDRs were estimated by identifying differential branchpoint usage between technical replicates. *P*-values were estimated by comparing the frequencies of differential branchpoint usage detected between tissues and between technical replicates (two-sided test for difference in proportion). (*E*) The fraction of constitutive introns exhibiting tissue-specific branchpoint usage within the GTEx data set. (*Left* panel) Introns binned by the total number of sequenced lariats across all 54 tissues sampled by the GTEx project. (*Right* panel) Introns binned by the mean number of sequenced lariats per tissue. Error bars indicate 95% confidence intervals estimated with a proportion test.

introns with two or more branchpoints, binned each intron according to the total number of sequenced lariats across all tissues, and tested whether each intron exhibited tissue-specific differences in branchpoint usage.

Our power analyses suggested that most branchpoints within constitutive introns are used relatively frequently within one or more tissues and that a majority of constitutive introns undergoes tissue-specific branchpoint usage (Fig. 5E). After binning introns by the total number of sequenced lariats, we found that ~87% of branchpoints within the highest-coverage bin (total of ≥250 sequenced lariats over all tissues) were used at rates of ≥10% in one or more tissues. Fifty-eight percent of constitutive introns within this highest-coverage bin exhibited tissue-specific branchpoint usage. We obtained even more striking results after binning introns by the mean number of sequenced lariats per tissue. This method more accurately controlled for how variable sequencing depth affected our power to quantify branchpoint usage. Approximately 96% of branchpoints within the highest-coverage bin (mean of ≥10 sequenced lariats per tissue) were used at rates of ≥10% in one or more tissues, and 81% of constitutive introns in this coverage bin exhibited tissue-specific branchpoint usage. We conclude that even constitutive introns commonly undergo tissue-specific branchpoint usage.

Discussion

In addition to providing a comprehensive genome-wide branchpoint annotation, our study has several important implications for future studies of splicing mechanisms and regulation. First, our finding that most introns have

multiple branchpoints suggests that any perturbation of branchpoint recognition may have unexpectedly profound consequences for global splicing. Branchpoint multiplicity may be particularly important in the context of cancer-associated mutations that alter normal splicing mechanisms and regulation (Dvinge et al. 2016). Second, our study demonstrates that branchpoint selection is unexpectedly complex in healthy tissues, even for constitutive introns. The striking tissue-specific variability in branchpoint usage that we observed suggests that introns are recognized in mechanistically distinct ways in different cell types.

The discovery of recurrent cancer-associated mutations affecting the splicing factor SF3B1 created intense interest in understanding how these mutations might alter normal splicing (Papaemmanuil et al. 2011; Quesada et al. 2011; Wang et al. 2011; Yoshida et al. 2011). SF3B1 is a core component of U2 snRNP that binds pre-mRNA near the branchpoint. While the mechanistic consequences of *SF3B1* mutations have not been fully elucidated, several studies have demonstrated that these lesions are associated with abnormal 3′ splice site recognition, including usage of cryptic 3′ splice sites and alternate branchpoints (Darman et al. 2015; DeBoever et al. 2015; Alsafadi et al. 2016). However, relatively few splicing changes have been identified to date in *SF3B1* mutant cells (Obeng et al. 2016). Given the unexpected branchpoint multiplicity and tissue-specific regulation revealed by our study, we speculate that *SF3B1* mutations might have more profound and pervasive consequences for global splicing than is currently recognized.

While our study highlights the complexity of recognizing even constitutive introns, the mechanistic origins of tissue-specific branchpoint usage remain mysterious. For example, it is unclear why different sets of branchpoints underlie *HBB* splicing in blood versus metastatic prostate cancer samples. This tissue specificity is not readily explained by somatic mutations, as the analyzed metastatic prostate cancer samples were not enriched for recurrent splicing factor mutations (Robinson et al. 2015). Interestingly, within each of the two distinct sets of branchpoints, the branchpoint with the best match to the U2 snRNA was dominant (Fig. 5A). Binding of *trans*-acting factors to the *HBB* intron might prevent branchpoint recognition by physical occlusion, and competition between nonoccluded branchpoints might govern subsequent patterns of branchpoint selection in a given tissue. Even a single splicing factor can occlude multiple branchpoints; for example, CELF2 can bind sites flanking a branchpoint cluster to simultaneously prevent usage of any branchpoint in the cluster (Dembowski and Grabowski 2009). Further studies are required to test whether this or other mechanisms enforce the tissue specificity of branchpoint usage within *HBB*, *VASP*, *SRSF3*, and other genes.

What are the functional consequences of branchpoint multiplicity? We speculate that having multiple branchpoints might confer both fitness advantages and expanded regulatory potential to introns. First, branchpoint multiplicity may confer mechanistic robustness. Having multiple branchpoints may render introns resilient to otherwise deleterious transcriptional errors, somatic mutations, or genetic variation. Second, branchpoint multiplicity may facilitate splicing regulation by rendering splice site recognition more plastic. For example, an intron with multiple branchpoints could be regulated by the intronic binding of tissue-specific splicing factors that promote or repress individual branchpoints. Such an intron might have more inherent regulatory potential than would an intron with a single branchpoint. This hypothesis is supported by our finding that introns associated with the highly regulated SR poison exons are rich with branchpoints. Third, branchpoint multiplicity may enable regulated retention of introns, including ostensibly constitutive introns. A majority of human introns, most of which are not associated with alternative splice site or exon usage, exhibits detectable intron retention in specific healthy and/or cancerous cell types (Braunschweig et al. 2014; Dvinge and Bradley 2015). With rare exceptions, the mechanistic origins of intron retention are not understood. However, our observation that the most frequently retained introns have more branchpoints than do other introns, including nonadenine and 3′ splice site-distal branchpoints, strongly suggests that branchpoint selection is an important contributor to regulated intron retention. While our understanding of branchpoint selection remains incomplete, it is clear that the branchpoint plays a more important regulatory role in both constitutive and alternative splicing than is generally recognized.

Materials and methods

Genome annotations

We generated a genome annotation by merging the UCSC knownGene (Meyer et al. 2013), Ensembl 71 (Flicek et al. 2013), and MISO version 2.0 (Katz et al. 2010) annotations for the UCSC hg19 (GRCh37) genome assembly. We created an expanded intron annotation for subsequent branchpoint mapping by enumerating all possible combinations of annotated 5′ and 3′ splice sites within each gene.

Gene expression and alternative splicing analysis

We estimated gene expression and alternative splicing across the 16 tissues in the Body Map 2.0 database as described previously (Dvinge et al. 2014). Briefly, we first mapped all reads to the transcriptome with RSEM (RNA-seq by expectation maximization) version 1.2.4 (Li and Dewey 2011), which produces gene-level expression estimates. We modified RSEM to invoke Bowtie (Langmead et al. 2009) with the option "-v 2." We then mapped remaining unaligned reads to the genome and the splice junction database described above (equivalent to the expanded intron annotation) with TopHat version 2.0.8b (Trapnell et al. 2009). We merged the read alignments produced by RSEM and TopHat and used those as input to MISO with its version 2.0 annotation (Katz et al. 2010) to quantify isoform expression.

Branchpoint detection algorithm

Our branchpoint detection algorithm was based on the split-read alignment strategy used in Mercer et al. (2015).

Prefilter reads First, filter out reads with >5% Ns or other ambiguous characters. Next, sequentially invoke Bowtie2 as follows for the transcriptome and genome: bowtie2 -x <index file for transcriptome or genome> --end-to-end --sensitive --score-min L,0,-0.24 -k 1 --n-ceil L,0,0.05 -U <FASTQ file of reads>. Finally, discard the aligned reads and use the unaligned reads as input for the next step.

Map 5′ splice site sequences to reads First, build a Bowtie index for a FASTA file of prefiltered reads. Next, build a FASTA file holding 5′ splice site sequences (the first 20 nt of each intron or, alternately, the 20 nt downstream from each 5′ splice site, including the 5′ splice site itself). Finally, map 5′ splice site sequences to reads as follows: bowtie2 -x <index file for reads> --end-to-end --sensitive --k 10000 --no-unal -f -U <FASTA file of 5′ splice site sequences>.

Restrict to reads aligned to a 5′ splice site and trim reads First, restrict to alignments between 5′ splice site sequences and reads with no mismatches and no indels. Second, restrict to reads that align to a single 5′ splice site. Third, trim off the portion of each read starting at the 5′ splice site alignment and continuing to the end of the read. Finally, restrict to trimmed reads of ≥20-nt length.

Map trimmed reads to 3′ splice site sequences First, build a FASTA file holding the 3′ splice site sequences (the last 250 nt of each intron or, alternately, the 250 nt upstream of each 3′ splice site, including the 3′ splice site itself). Next, build a Bowtie index for these sequences. Finally, map trimmed reads to 3′ splice sites as follows: bowtie2 -x <index file for 3′ splice sites> --end-to-end --sensitive -k 10 --no-unal -f -U <FASTA file of trimmed reads>.

Infer branchpoint positions from split-read alignments First, restrict to trimmed read alignments with five or fewer mismatches, ≤10% mismatch rate, and at most a single indel of ≤3-nt length in the 3′ splice site-aligning portion of the read. Second, restrict to alignments that score as well as the best-scoring alignment for each read (e.g., remove lower-scoring alignments). Third, restrict to reads with inverted alignments (e.g., where the "left" half of the read aligns near the 3′ splice site, while the "right" half of the read aligns to the 5′ splice site). Fourth, restrict to reads for which the 5′ and 3′ splice site-aligning portions of the read map to splice sites within a single gene. Fifth, compute the branchpoint position as the last nucleotide of the trimmed read alignment. Sixth, restrict to reads with a mismatch at the inferred branchpoint position. Finally, assemble a final set of 5′ splice site–branchpoint pairs.

Assign confidence levels to each 5′ splice site–branchpoint pair First, for each identified 5′ splice site–branchpoint pair, extract these sequences and identify nonunique sequences as follows: (1) 5′ splice site sequence (25 nt of sequence downstream from the 5′ splice site, including the 5′ splice site itself; test whether each sequence aligns to more than one location in the genome with no mismatches or gaps), (2) upstream branchpoint sequence (25 nt of sequence upstream of the branchpoint, including the branchpoint itself; test whether each sequence aligns to more than one location in the genome with no mismatches or gaps), and (3) lariat sequence (concatenation of the branchpoint and 5′ splice site sequence; test whether each sequence aligns to the transcriptome or genome with two or fewer mismatches, ≤5% mismatches, and no gaps). Next, assign confidence levels to each 5′ splice site–branchpoint pair based on the "hits" (branchpoint-spanning reads used to infer the branchpoint location) as follows: (1) low (one or more hits with mismatch at the branchpoint and ≥5% of hits with mismatches at the branchpoint), (2) moderate (one or more hits with mismatch at the branchpoint and no other mismatches or indels in the 3′ splice site region of the read and ≥5% of hits with mismatches at the branchpoint and no other mismatches or indels in the 3′ splice site region of the read), and (3) high (one or more hits with mismatch at the branchpoint and no other mismatches or indels in the 3′ splice site region of the read, ≥5% of hits with mismatches at the branchpoint and no other mismatches or indels in the 3′ splice site region of the read, and unique 5′ splice site, upstream branchpoint, and lariat sequences).

Branchpoint sequence analysis

Branchpoint sequence analysis was performed within the R programming environment. All analyses relied on Bioconductor tools, including the AnnotationHub, BSgenome, Genomic Alignments, GenomicFeatures, and GenomicRanges packages (Lawrence et al. 2013; Huber et al. 2015). All plots and figures were generated with the dplyr (http://CRAN.R-project.org/package=dplyr) and ggplot2 (Wickham 2009) packages.

Introns were classified as U2 or U12 type by computing the best match between the 5′ splice site of the intron and position weight matrices (PWMs) representing consensus U2- or U12-type 5′ splice sites (Sheth et al. 2006). The binding energy between a given branchpoint context (excluding the branchpoint itself, which appears as a bulge) and the U2 snRNA sequence AUGAUGUG was computed with ViennaRNA (Lorenz et al. 2011).

Several statistics were recomputed in order to create Table 1. The A frequency at branchpoints reported by Gao et al. (2008) was recomputed in order to be consistent with other studies. Gao et al. (2008) reported this statistic using all sequenced lariats rather than distinct called branchpoints. The A frequency was similarly recomputed for Taggart et al. (2012, 2017). For Taggart et al. (2017), the "high-confidence" set of branchpoints was used. This set of branchpoints was obtained by collapsing branchpoint calls (branchpoints in Supplemental Table S2 of Taggert et al. 2017 without motif model values of "template_switching" or "circle") (AJ Taggart and WG Fairbrother, pers. comm.).

Quantification of branchpoint usage across tissues

Branchpoint usage was computed across the 54 tissues represented in the GTEx data set (Melé et al. 2015) as follows. All reads were prefiltered by aligning them to the transcriptome and genome and then discarding aligned reads (as for the branchpoint discovery algorithm). Unaligned reads were collated across individuals for each tissue and aligned to a database of lariat sequences. The lariat sequence database consisted of sequences spanning the branchpoint itself, with the length of flanking sequence upstream of and downstream from the branchpoint chosen such that an aligned read must have at least 10 nt aligned to either side of the branchpoint. The database of lariat sequences is therefore dependent on the query read length. Unaligned reads were mapped to the lariat database with Bowtie2 (Langmead and Salzberg 2012). Alignments were restricted to those with three or fewer mismatches and one or fewer indel of ≤3-nt length. Reads were permitted to align to up to 25 different lariat sequences; however, multimapping reads were downweighted proportional to the number of lariats to which they aligned. Usage of a given branchpoint was then estimated as the number of reads supporting usage of that branchpoint divided by the total number of reads supporting usage of the 5′ splice site that was associated with that branchpoint. Usage of any branchpoint therefore must fall within the interval 0%–100%. This is analogous to the Ψ value commonly used for estimating usage of alternatively spliced sequence (Wang et al. 2008). For each branchpoint, a minimum of 20 reads per tissue was required in order to estimate branchpoint usage; if

<20 reads were available, then that data point was not subjected to further analysis.

An intron was said to exhibit tissue-specific branchpoint usage if we observed tissue-specific differences in branchpoint usage of ≥10% (absolute, rather than relative, value) with a *P*-value of ≤0.05 by a two-sided proportion test.

RNA extraction and cDNA synthesis

K562 cells were lysed using TRIzol (Thermo Fisher Scientific). K562 total RNA was isolated from the cell lysate according to the manufacturer's protocol and cleaned using the Qiagen RNeasy minikit with DNase treatment (Qiagen, RNase-free DNase set). Total RNA from human peripheral blood mononuclear cells, cerebella, testes, spleens, and fetal spleens were purchased from Takara Bio. cDNA was synthesized from the total RNA using random hexamer priming and the SuperScript III first strand synthesis system (Thermo Fisher Scientific).

Direct lariat sequencing and analysis

Primers (Integrated DNA Technologies) were designed to amplify the branchpoint–5′ splice site junction of a specific lariat via nested PCR as illustrated in Supplemental Figure S3A. A first round of gradient PCR (30 cycles) was performed with Phusion high-fidelity DNA polymerase (Thermo Fisher Scientific) using the "outer" primer set and the K562 cDNA as a template. The annealing temperatures used were in the range of Tm ±3°C. The reactions were pooled together, cleaned, and concentrated using the QIAquick PCR purification kit (Qiagen). The concentrated DNA served as the template for the second round of gradient PCR (30 cycles) using the "inner" primer set and an annealing temperature range analogous to that used for the first round of PCR. The reactions were combined and subjected to 2% agarose gel electrophoresis. Bands of sizes consistent with lariat amplification were excised, and DNA was extracted using the MinElute gel extraction kit (Qiagen). Purified DNA fragments were cloned into the pCR-Blunt II-TOPO vector (Thermo Fisher Scientific) and transformed into TOP10 chemically competent *Escherichia coli* (Thermo Fisher Scientific) using the ZeroBlunt TOPO PCR cloning kit (Thermo Fisher Scientific). The transformants were plated on 50 µg/mL LB + kanamycin plates, and random colonies were selected for Sanger sequencing (Genewiz) after growth. Inner and outer primer sets are listed in Supplemental Table S2.

Branchpoints were annotated using Sanger-sequenced amplicons with the algorithm outlined in Supplemental Figure S3B. Only reads that contained a mismatch or small insertion at the inferred branchpoint position were used to identify novel branchpoints. Reads with no lesion or those containing a deletion at the inferred branchpoint were used only to experimentally confirm branchpoints annotated via RNA-seq analysis, not to annotate new branchpoints. Tissue-specific branchpoint usage was quantified and analyzed analogously to the method described above for the GTEx data set.

Acknowledgments

J.M.B.P. was supported by the ARCS Foundation. R.K.B. is a Scholar of The Leukemia and Lymphoma Society (1344-18) and was supported by the Edward P. Evans Foundation, National Institutes of Health (NIH)/National Institute of Diabetes and Digestive and Kidney Diseases (R01 DK103854), and NIH/National Heart, Lung, and Blood Institute (NHLBI) (R01 HL128239). The results published here are based in part on data generated by The Cancer Genome Atlas Research Network (http://cancergenome.nih.gov). The Genotype-Tissue Expression (GTEx) Project was supported by the Common Fund of the Office of the Director of the NIH and by the National Cancer Institute, National Human Genome Research Institute, NHLBI, National Institute on Drug Abuse, National Institute of Mental Health, and National Institute of Neurological Disorders and Stroke. The GTEx data used for the analyses described in this manuscript were obtained from the Database of Genotypes and Phenotypes (dbGaP; accession no. phs000424.v6.p1) on February 8, 2017.

Author contributions: J.M.B.P. and R.K.B. performed the experiments, analyzed the data, and wrote the paper.

References

Alsafadi S, Houy A, Battistella A, Popova T, Wassef M, Henry E, Tirode F, Constantinou A, Piperno-Neumann S, Roman-Roman S, et al. 2016. Cancer-associated SF3B1 mutations affect alternative splicing by promoting alternative branchpoint usage. *Nat Commun* **7:** 10615.

Bradley RK, Merkin J, Lambert NJ, Burge CB. 2012. Alternative splicing of RNA triplets is often regulated and accelerates proteome evolution. *PLoS Biol* **10:** e1001229.

Braunschweig U, Barbosa-Morais NL, Pan Q, Nachman EN, Alipanahi B, Gonatopoulos-Pournatzis T, Frey B, Irimia M, Blencowe BJ. 2014. Widespread intron retention in mammals functionally tunes transcriptomes. *Genome Res* **24:** 1774–1786.

Corvelo A, Hallegger M, Smith CWJ, Eyras E. 2010. Genome-wide association between branch point properties and alternative splicing. *PLoS Comput Biol* **6:** e1001016.

Darman RB, Seiler M, Agrawal AA, Lim KH, Peng S, Aird D, Bailey SL, Bhavsar EB, Chan B, Colla S, et al. 2015. Cancer-associated SF3B1 hotspot mutations induce cryptic 3′ splice site selection through use of a different branch point. *Cell Rep* **13:** 1033–1045.

DeBoever C, Ghia EM, Shepard PJ, Rassenti L, Barrett CL, Jepsen K, Jamieson CHM, Carson D, Kipps TJ, Frazer KA. 2015. Transcriptome sequencing reveals potential mechanism of cryptic 3′ splice site selection in SF3B1-mutated cancers. *PLoS Comput Biol* **11:** e1004105.

Dembowski JA, Grabowski PJ. 2009. The CUGBP2 splicing factor regulates an ensemble of branchpoints from perimeter binding sites with implications for autoregulation. *PLoS Genet* **5:** e1000595.

Dietrich RC, Peris MJ, Seyboldt AS, Padgett RA. 2001. Role of the 3′ splice site in U12-dependent intron splicing. *Mol Cell Biol* **21:** 1942–1952.

Dvinge H, Bradley RK. 2015. Widespread intron retention diversifies most cancer transcriptomes. *Genome Med* **7:** 45.

Dvinge H, Ries RE, Ilagan JO, Stirewalt DL, Meshinchi S, Bradley RK. 2014. Sample processing obscures cancer-specific alterations in leukemic transcriptomes. *Proc Natl Acad Sci* **111:** 16802–16807.

Dvinge H, Kim E, Abdel-Wahab O, Bradley RK. 2016. RNA splicing factors as oncoproteins and tumour suppressors. *Nat Rev Cancer* **16:** 413–430.

Flicek P, Ahmed I, Amode MR, Barrell D, Beal K, Brent S, Carvalho-Silva D, Clapham P, Coates G, Fairley S, et al. 2013. Ensembl 2013. *Nucleic Acids Res* **41:** D48–D55.

Fu X-D, Ares M. 2014. Context-dependent control of alternative splicing by RNA-binding proteins. *Nat Rev Genet* **15:** 689–701.

Gao K, Masuda A, Matsuura T, Ohno K. 2008. Human branch point consensus sequence is yUnAy. *Nucleic Acids Res* **36:** 2257–2267.

Gattoni R, Schmitt P, Stevenin J. 1988. In vitro splicing of adenovirus E1A transcripts: characterization of novel reactions and of multiple branch points abnormally far from the 3′ splice site. *Nucleic Acids Res* **16:** 2389–2409.

Hallegger M, Sobala A, Smith CWJ. 2010. Four exons of the serotonin receptor 4 gene are associated with multiple distant branch points. *RNA* **16:** 839–851.

Hartmuth K, Barta A. 1988. Unusual branch point selection in processing of human growth hormone pre-mRNA. *Mol Cell Biol* **8:** 2011–2020.

Helfman DM, Ricci WM. 1989. Branch point selection in alternative splicing of tropomyosin pre-mRNAs. *Nucleic Acids Res* **17:** 5633–5650.

Hornig H, Aebi M, Weissmann C. 1986. Effect of mutations at the lariat branch acceptor site on β-globin pre-mRNA splicing in vitro. *Nature* **324:** 589–591.

Huber W, Carey VJ, Gentleman R, Anders S, Carlson M, Carvalho BS, Bravo HC, Davis S, Gatto L, Girke T, et al. 2015. Orchestrating high-throughput genomic analysis with Bioconductor. *Nat Methods* **12:** 115–121.

Katz Y, Wang ET, Airoldi EM, Burge CB. 2010. Analysis and design of RNA sequencing experiments for identifying isoform regulation. *Nat Methods* **7:** 1009–1015.

Kol G, Lev-Maor G, Ast G. 2005. Human-mouse comparative analysis reveals that branch-site plasticity contributes to splicing regulation. *Hum Mol Genet* **14:** 1559–1568.

Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9:** 357–359.

Langmead B, Trapnell C, Pop M, Salzberg SL. 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* **10:** R25.

Lareau LF, Inada M, Green RE, Wengrod JC, Brenner SE. 2007. Unproductive splicing of SR genes associated with highly conserved and ultraconserved DNA elements. *Nature* **446:** 926–929.

Lawrence M, Huber W, Pagès H, Aboyoun P, Carlson M, Gentleman R, Morgan MT, Carey VJ. 2013. Software for computing and annotating genomic ranges. *PLoS Comput Biol* **9:** e1003118.

Li B, Dewey CN. 2011. RSEM: accurate transcript quantification from RNA-seq data with or without a reference genome. *BMC Bioinformatics* **12:** 323.

Lorenz R, Bernhart SH, Höner Zu Siederdissen C, Tafer H, Flamm C, Stadler PF, Hofacker IL. 2011. ViennaRNA package 2.0. *Algorithms Mol Biol* **6:** 26.

Melé M, Ferreira PG, Reverter F, DeLuca DS, Monlong J, Sammeth M, Young TR, Goldmann JM, Pervouchine DD, Sullivan TJ, et al. 2015. Human genomics. The human transcriptome across tissues and individuals. *Science* **348:** 660–665.

Mercer TR, Clark MB, Andersen SB, Brunck ME, Haerty W, Crawford J, Taft RJ, Nielsen LK, Dinger ME, Mattick JS. 2015. Genome-wide discovery of human splicing branchpoints. *Genome Res* **25:** 290–303.

Meyer LR, Zweig AS, Hinrichs AS, Karolchik D, Kuhn RM, Wong M, Sloan CA, Rosenbloom KR, Roe G, Rhead B, et al. 2013. The UCSC Genome Browser database: extensions and updates 2013. *Nucleic Acids Res* **41:** D64–D69.

Mullen MP, Smith CW, Patton JG, Nadal-Ginard B. 1991. α-Tropomyosin mutually exclusive exon selection: competition between branchpoint/polypyrimidine tracts determines default exon choice. *Genes Dev* **5:** 642–655.

Ni JZ, Grate L, Donohue JP, Preston C, Nobida N, O'Brien G, Shiue L, Clark TA, Blume JE, Ares M. 2007. Ultraconserved elements are associated with homeostatic control of splicing regulators by alternative splicing and nonsense-mediated decay. *Genes Dev* **21:** 708–718.

Noble JC, Pan ZQ, Prives C, Manley JL. 1987. Splicing of SV40 early pre-mRNA to large T and small t mRNAs utilizes different patterns of lariat branch sites. *Cell* **50:** 227–236.

Noble JC, Prives C, Manley JL. 1988. Alternative splicing of SV40 early pre-mRNA is determined by branch site selection. *Genes Dev* **2:** 1460–1475.

Obeng EA, Chappell RJ, Seiler M, Chen MC, Campagna DR, Schmidt PJ, Schneider RK, Lord AM, Wang L, Gambe RG, et al. 2016. Physiologic expression of Sf3b1(K700E) causes impaired erythropoiesis, aberrant splicing, and sensitivity to therapeutic spliceosome modulation. *Cancer Cell* **30:** 404–417.

O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, Rajput B, Robbertse B, Smith-White B, Ako-Adjei D, et al. 2016. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res* **44:** D733–D745.

Papaemmanuil E, Cazzola M, Boultwood J, Malcovati L, Vyas P, Bowen D, Pellagatti A, Wainscoat JS, Hellstrom-Lindberg E, Gambacorti-Passerini C, et al. 2011. Somatic SF3B1 mutation in myelodysplasia with ring sideroblasts. *N Engl J Med* **365:** 1384–1395.

Quesada V, Conde L, Villamor N, Ordóñez GR, Jares P, Bassaganyas L, Ramsay AJ, Beà S, Pinyol M, Martínez-Trillos A, et al. 2011. Exome sequencing identifies recurrent mutations of the splicing factor SF3B1 gene in chronic lymphocytic leukemia. *Nat Genet* **44:** 47–52.

Reed R, Maniatis T. 1988. The role of the mammalian branchpoint sequence in pre-mRNA splicing. *Genes Dev* **2:** 1268–1276.

Robinson D, Van Allen EM, Wu YM, Schultz N, Lonigro RJ, Mosquera JM, Montgomery B, Taplin ME, Pritchard CC, Attard G, et al. 2015. Integrative clinical genomics of advanced prostate cancer. *Cell* **161:** 1215–1228.

Ruskin B, Krainer AR, Maniatis T, Green MR. 1984. Excision of an intact intron as a novel lariat structure during pre-mRNA splicing in vitro. *Cell* **38:** 317–331.

Ruskin B, Greene JM, Green MR. 1985. Cryptic branch point activation allows accurate in vitro splicing of human β-globin intron mutants. *Cell* **41:** 833–844.

Scotti MM, Swanson MS. 2016. RNA mis-splicing in disease. *Nat Rev Genet* **17:** 19–32.

Sheth N, Roca X, Hastings ML, Roeder T, Krainer AR, Sachidanandam R. 2006. Comprehensive splice-site analysis using comparative genomics. *Nucleic Acids Res* **34:** 3955–3967.

Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, Clawson H, Spieth J, Hillier LW, Richards S, et al. 2005. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res* **15:** 1034–1050.

Smith CW, Chu TT, Nadal-Ginard B. 1993. Scanning and competition between AGs are involved in 3′ splice site selection in mammalian introns. *Mol Cell Biol* **13:** 4939–4952.

Southby J, Gooding C, Smith CW. 1999. Polypyrimidine tract binding protein functions as a repressor to regulate alternative splicing of α-actinin mutally exclusive exons. *Mol Cell Biol* **19:** 2699–2711.

Taggart AJ, Desimone AM, Shih JS, Filloux ME, Fairbrother WG. 2012. Large-scale mapping of branchpoints in human pre-mRNA transcripts in vivo. *Nat Struct Mol Biol* **19:** 719–721.

Taggart AJ, Lin C-L, Shrestha B, Heintzelman C, Kim S, Fairbrother WG. 2017. Large-scale analysis of branchpoint usage across species and cell lines. *Genome Res* **27:** 639–649.

Trapnell C, Pachter L, Salzberg SL. 2009. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* **25:** 1105–1111.

Vogel J, Hess WR, Börner T. 1997. Precise branch point mapping and quantification of splicing intermediates. *Nucleic Acids Res* **25:** 2030–2031.

Wahl MC, Will CL, Lührmann R. 2009. The spliceosome: design principles of a dynamic RNP machine. *Cell* **136:** 701–718.

Wang ET, Sandberg R, Luo S, Khrebtukova I, Zhang L, Mayr C, Kingsmore SF, Schroth GP, Burge CB. 2008. Alternative isoform regulation in human tissue transcriptomes. *Nature* **456:** 470–476.

Wang L, Lawrence MS, Wan Y, Stojanov P, Sougnez C, Stevenson K, Werner L, Sivachenko A, DeLuca DS, Zhang L, et al. 2011. SF3B1 and other novel cancer genes in chronic lymphocytic leukemia. *N Engl J Med* **365:** 2497–2506.

Wickham H. 2009. *ggplot2: elegant graphics for data analysis*. Springer, New York.

Yoshida K, Sanada M, Shiraishi Y, Nowak D, Nagata Y, Yamamoto R, Sato Y, Sato-Otsubo A, Kon A, Nagasaki M, et al. 2011. Frequent pathway mutations of splicing machinery in myelodysplasia. *Nature* **478:** 64–69.

Zhuang YA, Goldstein AM, Weiner AM. 1989. UACUAAC is the preferred branch site for mammalian mRNA splicing. *Proc Natl Acad Sci* **86:** 2752–2756.

Most human introns are recognized via multiple and tissue-specific branchpoints

Jose Mario Bello Pineda and Robert K. Bradley

Genes Dev. published online April 17, 2018
Access the most recent version at doi:10.1101/gad.312058.118