Genome-wide discovery of human splicing branchpoints

Tim R. Mercer,[1,2,8] Michael B. Clark,[1,3,8] Stacey B. Andersen,[4] Marion E. Brunck,[4] Wilfried Haerty,[3] Joanna Crawford,[5] Ryan J. Taft,[5,6,7] Lars K. Nielsen,[4] Marcel E. Dinger,[1,2] and John S. Mattick[1,2]

[1]Garvan Institute of Medical Research, Sydney, New South Wales 2010, Australia; [2]St. Vincent's Clinical School, Faculty of Medicine, UNSW Australia, Sydney, New South Wales 2052, Australia; [3]MRC Functional Genomics Unit, Department of Physiology, Anatomy, and Genetics, University of Oxford, Oxford OX1 3PT, United Kingdom; [4]Australian Institute for Bioengineering and Nanotechnology, The University of Queensland, Brisbane, Queensland 4072, Australia; [5]Institute for Molecular Bioscience, The University of Queensland, Brisbane, Queensland 4072, Australia; [6]Illumina, Inc., San Diego, California 92122, USA; [7]School of Medicine and Health Services, Department of Integrated Systems Biology and Department of Pediatrics, George Washington University, Washington DC 20037, USA

During the splicing reaction, the 5′ intron end is joined to the branchpoint nucleotide, selecting the next exon to incorporate into the mature RNA and forming an intron lariat, which is excised. Despite a critical role in gene splicing, the locations and features of human splicing branchpoints are largely unknown. We use exoribonuclease digestion and targeted RNA-sequencing to enrich for sequences that traverse the lariat junction and, by split and inverted alignment, reveal the branchpoint. We identify 59,359 high-confidence human branchpoints in >10,000 genes, providing a first map of splicing branchpoints in the human genome. Branchpoints are predominantly adenosine, highly conserved, and closely distributed to the 3′ splice site. Analysis of human branchpoints reveals numerous novel features, including distinct features of branchpoints for alternatively spliced exons and a family of conserved sequence motifs overlapping branchpoints we term B-boxes, which exhibit maximal nucleotide diversity while maintaining interactions with the keto-rich U2 snRNA. Different B-box motifs exhibit divergent usage in vertebrate lineages and associate with other splicing elements and distinct intron–exon architectures, suggesting integration within a broader regulatory splicing code. Lastly, although branchpoints are refractory to common mutational processes and genetic variation, mutations occurring at branchpoint nucleotides are enriched for disease associations.

[Supplemental material is available for this article.]

The majority of human genes are spliced, a process whereby introns are removed from the nascent RNA and the remaining exonic sequence joined together into a mature RNA transcript. In addition, alternative splicing generates complex networks of isoforms from human gene loci and plays a major role in shaping the diversity of the transcriptome (Kapranov et al. 2005; Gerstein et al. 2007; Djebali et al. 2012).

Splicing occurs in the spliceosome, a large ribonucleoprotein complex that recognizes at least three genetic elements within each intron: the 5′ splice site (5′SS), the 3′ splice site (3′SS), and the branchpoint (Will and Lührmann 2011). *RNU2-1*, the U2 spliceosomal RNA (snRNA) base pairs to the sequence surrounding the unpaired branchpoint nucleotide, which then undergoes *trans*-esterification with the 5′ end of the intron to form a closed lariat structure. The spliceosome then scans for the downstream 3′ splice site, which undergoes a second *trans*-esterification reaction to join together the two exon ends and excise the intron lariat (Fig. 1; Smith et al. 1989; Will and Lührmann 2011).

Branchpoint selection by the spliceosome is an early step in the splicing reaction and one that subsequently defines the 3′ splice site and leads to inclusion of the downstream exon in the mature RNA (Hornig et al. 1986; Reed and Maniatis 1988; Smith et al. 1993). Mutations that abolish the branchpoint nucleotide can result in exon skipping and aberrant splicing and cause human disease (Padgett 2012; Singh and Cooper 2012). Despite such importance, only a few hundred human branchpoints have been identified, thereby preventing a detailed analysis of this basal splicing element (Gao et al. 2008; Taggart et al. 2012; Bitton et al. 2014).

Human branchpoints are difficult to predict by sequence alone due to the reported high degeneracy of the sequence around the branchpoint (Gao et al. 2008). Further complicating de novo branchpoint predictions are reports that some introns have multiple branchpoints, branchpoints distal from the 3′ splice site, or without the canonical adenine base at the branchpoint nucleotide (Gao et al. 2008; Taggart et al. 2012; Bitton et al. 2014). However, during the preparation of cDNA, reverse transcriptase can traverse the 5′SS/branchpoint junction, thereby generating a cDNA with the 5′ intronic sequence juxtaposed to the sequence immediately upstream of the branchpoint (Vogel et al. 1997). The split and inverted alignment of the resultant cDNA identifies not only the 5′ donor splice site but also the branchpoint nucleotide (Fig. 1; Taggart et al. 2012; Bitton et al. 2014), yet such events are encountered so rarely as to have hitherto prevented efficient and comprehensive genome scale annotation.

[8] **These authors contributed equally to this work.**

Corresponding author: j.mattick@garvan.org.au

Article published online before print. Article, supplemental material, and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.182899.114.

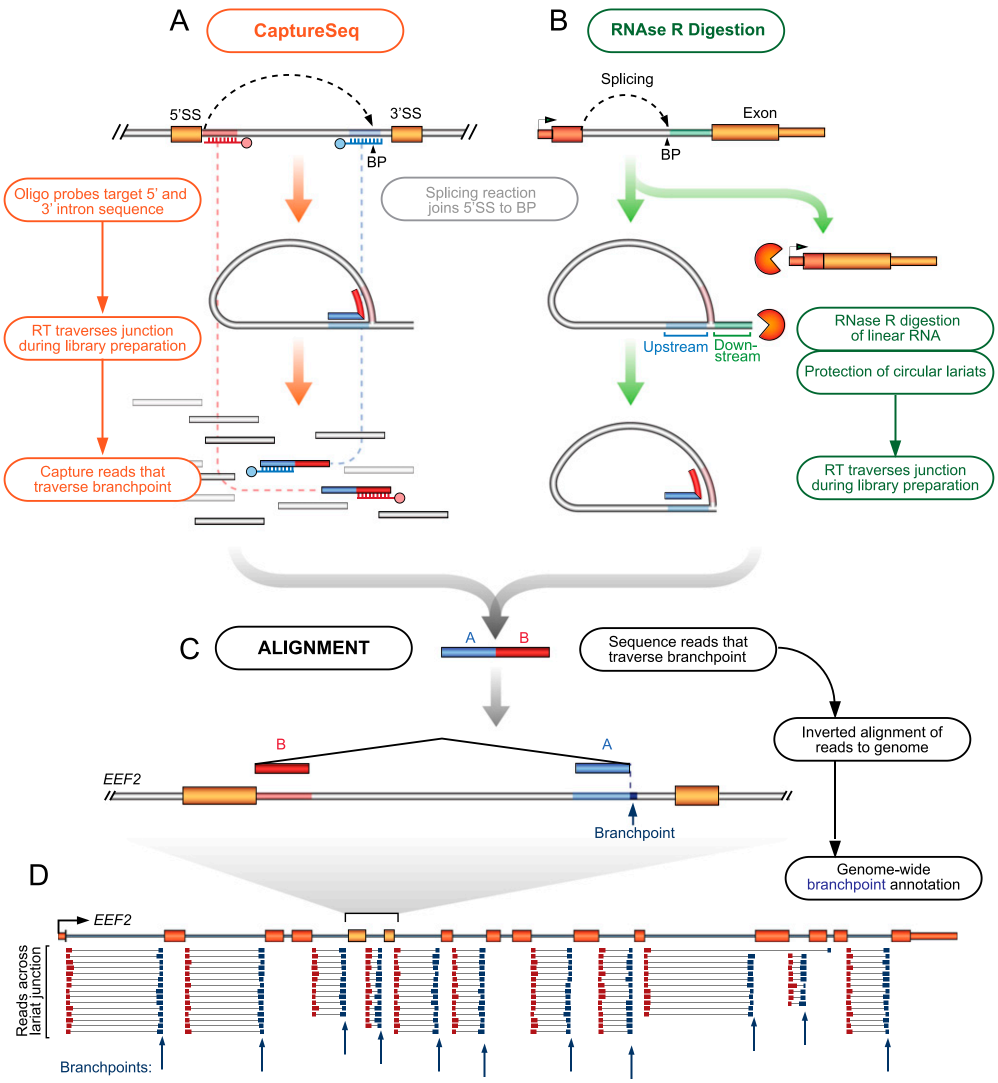
A
CaptureSeq
5'SS
3'SS
BP
Oligo probes target 5' and 3' intron sequence
Splicing reaction joins 5'SS to BP
RT traverses junction during library preparation
Capture reads that traverse branchpoint
B
RNAse R Digestion
Splicing
Exon
BP
Upstream
Down-stream
RNase R digestion of linear RNA
Protection of circular lariats
RT traverses junction during library preparation
C
ALIGNMENT
A
B
Sequence reads that traverse branchpoint
Inverted alignment of reads to genome
Genome-wide branchpoint annotation
EEF2
Branchpoint
D
EEF2
Reads across lariat junction
Branchpoints:

Figure 1. Identification of splicing branchpoints. Splicing joins the 5′ splice site to the branchpoint (BP) nucleotide near the 3′ end of the intron to form a lariat. Reverse transcriptase can traverse the branched 2′ to 5′ junction to generate informative sequencing reads (blue/red reads) that indicate the branchpoint location when aligned to the genome. We pursued two alternative strategies to enrich for sequenced reads containing branchpoints. (*A*) CaptureSeq (*left* orange pathway) uses labeled oligonucleotide probes to capture informative reads for targeted sequencing. (*B*) RNase R (*right* green pathway) digests linear mRNAs and selectively purifies circular RNAs including intron lariats. (*C*) Following sequencing, informative reads are split and inverted for alignment to the reference genome (black *lower* pathway), with 3′ termini of split alignment indicating the branchpoint nucleotide. (*D*) Example of informative alignments used to identify branchpoints (blue arrows) within the *EEF2* gene.

To enrich for such rare cDNAs that traverse the branchpoint junction, we have used two complementary approaches: purification of lariats by exoribonuclease digestion of linear RNAs; and capture of the branchpoint junction with oligonucleotide probes for targeted RNA sequencing. The sequencing and alignment of these approaches enables the first genome-wide discovery and analysis of human splicing branchpoints.

Results

Intron lariat sequencing and alignment

We recently developed RNA Capture Sequencing (CaptureSeq), a technique that focuses sequencing on targeted RNA transcripts (Mercer et al. 2012, 2014). This approach is ideal for achieving high sequencing coverage of transient intronic species to resolve branchpoint sites. We designed oligonucleotide probes targeting human intronic sequences immediately upstream of the 3′ splice site as well as 5′ intronic sequences that, following the 2′ *trans*-esterification reaction, lie immediately downstream from the branchpoint junction. By targeting lariat sequences both upstream of and downstream from putative branchpoint locations, we thereby enrich for cDNA transcripts that traverse the lariat junction (Fig. 1A; Supplemental Fig. 1A,B; Supplemental Data 1). We interrogated 238,576 introns (90.2% of annotated total) in this manner. We performed CaptureSeq similarly to previously described methods (Mercer et al. 2012) with total RNA harvested from three biological replicates of human K562 cells (see Methods). We achieved greater than 100-fold enrichment of targeted RNA as determined by qPCR and analysis of RNA spike-in controls, corresponding to ∼10.6-fold more reads (∼133 million reads) that align to introns relative to matched RNA-seq controls (Supplemental Fig. 1C,D).

As an alternative strategy, we also used RNase R digestion to selectively enrich for intron lariats. RNase R is a 3′ to 5′ exoribonuclease that digests linear RNAs but not circular lariats (Fig. 1B; Suzuki et al. 2006). We performed RNase R digestion in duplicate on two human cell-types (K562 and HeLa) along with matched mock-digested controls (see Methods; Supplemental Fig. 2A). We confirmed circular RNA protection using PCR targeting linear and circular isoforms of *MAN1A2* and *FBXW4* genes (Supplemental Fig. 2B,C; Salzman et al. 2012). Digested libraries were sequenced and aligned to the genome, indicating a global 9.6-fold global enrichment for introns lariats and commensurate depletion of exons (Supplemental Fig. 2D).

Informative sequenced reads that traverse the 5′SS/branchpoint junction require nonconventional alignment to the genome. From reads that do not otherwise align to the genome, we first filtered for reads containing 5′ intronic sequence before trimming the 5′ intronic sequence from the read and aligning the remaining sequence to introns at the same loci, using the 3′ termini of the read to indicate the branchpoint nucleotide (see Methods; Supplemental Fig. 3A). We applied this alignment strategy to both RNase R and CaptureSeq libraries, as well as to 234 publicly available conventional RNA sequencing libraries derived from a wide range of human cell types (Djebali et al. 2012). Using this approach, we aligned 532,405 informative sequenced reads that identified 88,748 branchpoints. The relative proportion of sense/antisense alignment estimated a 0.2% false alignment rate (see Methods). For comparison, we returned one informative read for each 7473 sequenced reads using CaptureSeq, one per 41,725 reads from RNase R digestion and sequencing, and one informative read per 146,617 reads sequenced with standard RNA-seq (Supplemental Fig. 3B). We also determined the contribution of each technique to the identification of high-confidence branchpoints (Supplemental Fig. 3E). Despite the large number of conventional RNA sequencing libraries we examined, the majority of unique branchpoints were discovered by CaptureSeq and RNase R, confirming their value for branchpoint identification. Lastly, we detected two biases within our branchpoint annotations. Given that sequence coverage of branch junctions does not reach saturation, highly expressed genes are better represented within branchpoint annotations; and although we identify branchpoints in introns larger than 100 kb, we observe an overall bias for branchpoints in short introns.

Validation of branchpoint annotations

Nucleotide misprocessing often occurs when reverse transcriptase traverses the noncanonical 2′ to 5′ linkage between the 5′ splice site nucleotide and branchpoint at the lariat junction, resulting in a mismatch error, microinsertion or deletion in the generated cDNA (Fig. 2A; Vogel et al. 1997; Gao et al. 2008). Sequenced reads with small insertions or deletions are present at 8673 branchpoints but are unable to precisely locate the branchpoint nucleotide (Supplemental Fig. 3C). However, single mismatch errors in the cDNA are diagnostic of the exact branchpoint nucleotide (Gao et al. 2008). We found 70.4% of sequenced reads (375,178 total) aligning to branchpoints contained mismatches at the branchpoint nucleotide, comprising an 392.1-fold higher error rate over background and confirming the location of 74.1% (59,359) of branchpoint nucleotides (Fig. 2B,C). Branchpoint annotations for each level of supporting evidence are provided (Fig. 2D; Supplemental Table 1; Supplemental Data 2).

The majority of branchpoints (75.2%) are supported by multiple reads, with coverage encompassing a 10^3-fold dynamic range (Supplemental Fig. 3D). This quantitative profile is closely correlated between biological replicates (Pearson's $r^2 = 0.92$) and provides a quantitative measure of branchpoint selection preference.

Detailed analysis of RNase R-digested alignment profiles also revealed the selective digestion of lariat regions downstream from the branchpoint, which corresponds to the exposed linear tail, whereas the regions upstream of the branchpoint corresponding to the closed circular lariat remain protected (Fig. 3A,D). We observed a global 3.7-fold depletion ($P = 0.029$, paired *t*-test) of intron lariats immediately downstream from branchpoint annotations (corresponding to the lariat tail) in RNase R-treated samples (with 26,758 branchpoints exhibiting greater than twofold downstream depletion) (Fig. 3B,C), lending further support to branchpoint locations identified by RNase R digestion. Intron-specific RT-PCR amplification followed by Sanger sequencing was also performed to independently confirm the location of 10 branchpoint nucleotides (Supplemental Fig. 3F).

Branchpoint features

This first large-scale experimental annotation of branchpoints afforded an opportunity for the global analysis of branchpoint features. We restricted our analysis to the 59,359 high-confidence branchpoints confirmed by mismatch sequencing errors that correspond to ∼17.4% of introns occurring within 10,773 genes (24.8% of total).

Branchpoint annotations exhibited several previously described features (Gao et al. 2008; Corvelo et al. 2010; Taggart et al. 2012), including a restricted distribution upstream of the 3′ splice site, with 90% of branchpoints occurring within 19 to 37 (median 25) nucleotides upstream (Fig. 4B). The majority of branchpoint nucleotides correspond to adenine (78.4%), with lower frequency selection of cytosine (8.4%), uracil (8.4%), and guanine (4.7%) nucleotides that have been shown to function, albeit with lower efficiency, as branchpoints (Fig. 4A; Reed and Maniatis 1988). Three noncanonical branchpoints were selected and confirmed by RT-PCR/Sanger sequencing (Supplemental Fig. 3F). The region downstream from the branchpoint exhibits a strong depletion of AG dinucleotides (4.0-fold depletion, with 77% of intervening regions containing no AG dinucleotides with the exception of the 3′ splice site), consistent with the spliceosomal scanning model of 3′ splice site recognition (Fig. 4D; Supplemental Fig. 4A; Smith et al. 1989).

Genome-wide maps also showed new branchpoint features. A large proportion of exons (32.4%) were associated with more than one branchpoint, indicating flexibility or redundancy in branchpoint selection (Supplemental Data 3). These multiple branchpoints are generally clustered in close proximity to each other and conform to a similarly tight distribution in relation to the 3′ splice site as for individual branchpoints (Fig. 4E), although exceptional distal branchpoints were identified (Supplemental Fig. 4B). Multiple branchpoints were not equally used, with the majority of exons (52.9%) having one or a small number of dominant branchpoints. Multiple branchpoints exhibit lower conservation and a lower preference for adenine (Supplemental Fig. 4C). Dense multiple branchpoint clusters are notably enriched for cytosine branchpoints that are conserved relative to surrounding sequence (Supplemental Fig. 4D). Cytosine branchpoints have been shown to be resistant to debranching (Hornig et al. 1986) and may comprise a distinct subset recently associated with stable lariat and long noncoding RNA formation (Zhang et al. 2013).

Alternative splicing can also be detected at the branchpoint level. Using our split/inverted read alignments to pair branchpoints with their partner 3′ and 5′ splice sites, we identify alternative splic-

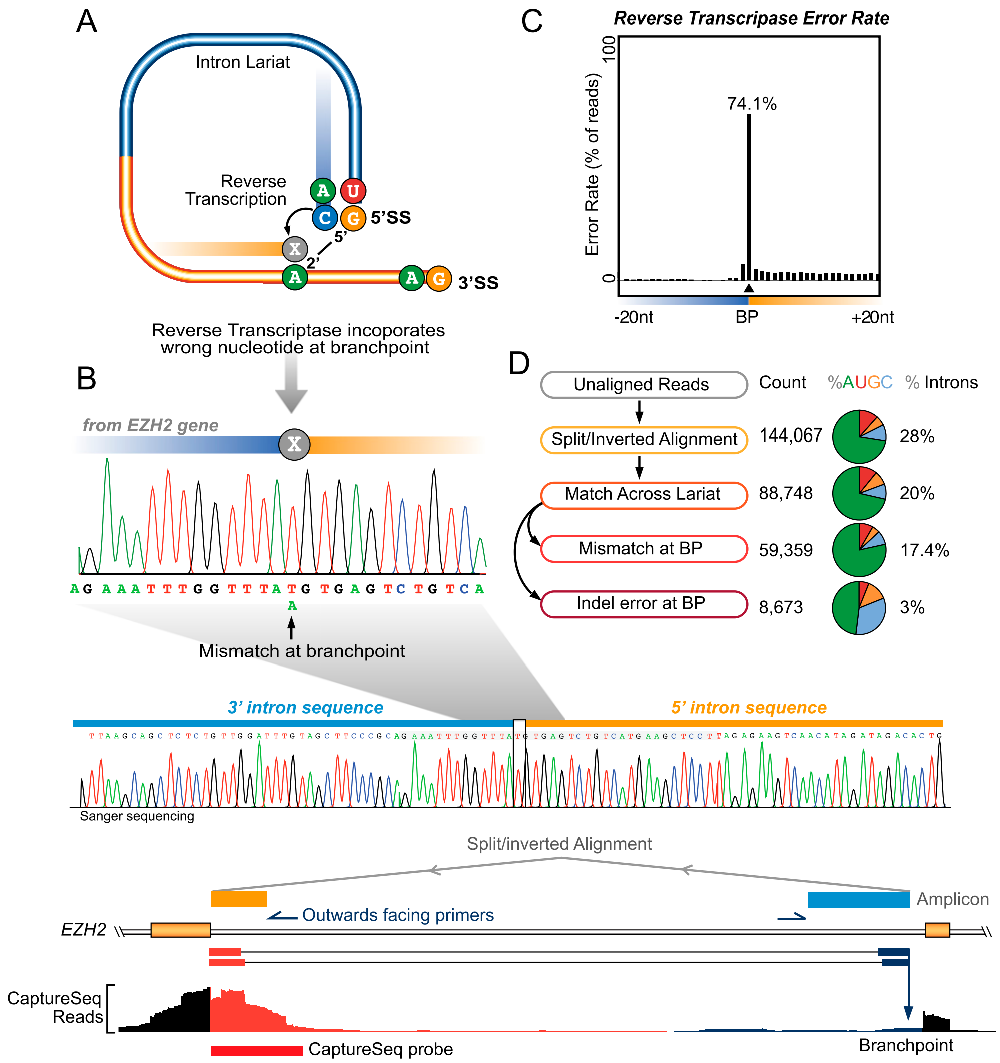

Figure 2. Validation of branchpoint annotations. (*A*) Schematic indicating incorporation of mismatch nucleotide (gray X) within cDNA (blue/orange bar) by reverse transcriptase when traversing the 2′ to 5′ branch junction. (*B*) UCSC Genome Browser view of branchpoints within *EZH2* gene (*lower* panel) showing capture of sequences at intron ends (histogram) and sequenced read alignments that indicate branchpoints (red/blue bar). RT-PCR using *outward* facing primers (blue) generates an amplicon that validates branchpoint annotations (blue/orange *upper* panel). Zoom of chromatograph (*top left*) shows sequence mismatch error (T) at the branchpoint nucleotide (boxed). (*C*) Histogram indicates rate of mismatch errors at the branchpoint nucleotide in sequenced reads that align across the lariat junction (*center*). (*D*) Summary of branchpoint annotations with varying levels of support: split/inverted alignment, exact match across lariat junction, match with mismatch, or insertion/deletion errors at branchpoints. At each level of support, the number of identified branchpoints, nucleotide composition of the branchpoint, and fraction of GENCODE introns with an annotated branchpoint are indicated.

ing events within 12.6% of alignments, whereby multiple intron 5′ termini are joined to a common branchpoint, or a common 5′ termini is joined to multiple branchpoints (Fig. 4F). Examples of differential branchpoint selection and usage during alternative splicing were also observed at the branchpoint level, with alternative 5′ splice sites exhibiting different preferences for multiple branchpoints that precede a single common exon (Supplemental Fig. 4E).

The vast majority of identified branchpoints are within protein-coding genes; however, we also identify 551 branchpoints from 255 long noncoding RNA (lncRNA) loci, including well-known lncRNAs such as *HOTAIR*, *XIST*, and a number of snoRNA-host lncRNAs.

Branchpoint identification can also lend support to the inclusion of associated exons and exon variants presently absent from comprehensive gene catalogs such as GENCODE (Harrow et al. 2012). Strict filtering identified a group of 16 exons, including a conserved 21-nt micro-exon in the *MAST2* gene (Supplemental Fig. 4F; Supplemental Table 2).

Branchpoint motifs

The nucleotides flanking branchpoints bind U2 snRNA and are highly conserved, suggesting the proximity of regulatory sequence elements (Fig. 4C). Extreme examples of such proximal conservation include branchpoints found within ultraconserved elements that govern the unproductive splicing of the *SR* gene family (Supplemental Fig. 5A; Lareau et al. 2007).

To identify *cis*-elements, we performed de novo motif identification around the branchpoint (Neph et al. 2012). This resolved a set of 5- to 6-nt sequence motifs that overlap ∼53% of all branchpoint annotations and can undergo base-pairing interactions with the U2 snRNA (Fig. 5A; Supplemental Fig. 5B). We term these B-box

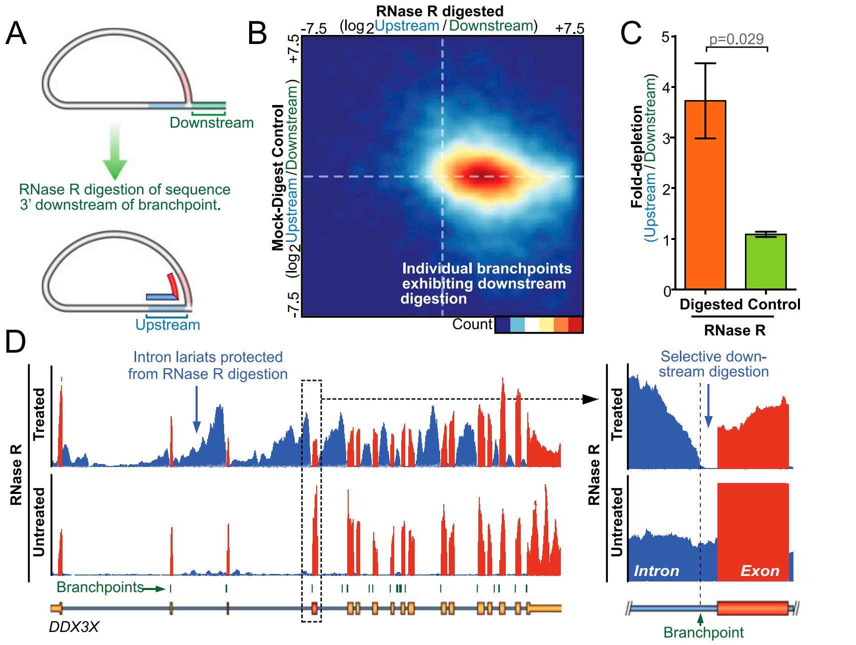

Figure 3. RNase R digestion of intron lariats. (*A*) Schematic showing digestion of the lariat tail (green) downstream from the branchpoint by RNase R, while the closed lariat protects sequences upstream of the branchpoint (blue). (*B*) Heatmap indicates the relative enrichment of sequences upstream of individual branchpoints within RNase R digested libraries (*x*-axis) compared to untreated control (*y*-axis) libraries. The majority of branchpoints show enrichment of upstream sequences after digestion (indicated by positive *horizontal* spread), whereas the clustering at zero on the *y*-axis indicates that sequences around individual branchpoints show no enrichment in mock-digested libraries. (*C*) Histogram indicates significant depletion of intronic sequences downstream from branchpoints (to the 3′SS) relative to the upstream region following RNase R digestion. Two human cell types, each with two biological replicates with matched untreated controls (paired *t*-test, $n = 4$, error bars SEM). (*D*) Genome Browser view showing enrichment of *DDX3X* intron lariats following RNase R digestion. Close detail indicates selective digestion of intronic sequence downstream from the identified branchpoint. In contrast, equivalent pre-mRNA is observed both up- and downstream from branchpoints in untreated libraries.

elements since they are enriched for B-nucleotides (C, G, and U) (Cornish-Bowden 1985) and depleted of adenosine nucleotides (with the exception of the branchpoint itself). B-box nucleotides base pair with the high density of keto-residues (G and U) within the U2 snRNA, and the abundance and conservation of different B-box families correlates with predicted U2 binding strength (Fig. 5B; Supplemental Fig. 5C). Due to RNA wobble-base pairing, the U2 snRNA keto-residues can interact with two possible nucleotides at the opposing position in the B-box element. This enables B-boxes to have greater informational diversity while maintaining their ability to bind U2 snRNAs (Fig. 5A,C). This diversity underpins the observed range of sequence families that when collectively analyzed have been previously considered degenerate (Gao et al. 2008). Similar enriched interactions between B- and keto-nucleotides have been observed for microRNA seed sequences (Nelson and Green 1989; Wang 2013).

B-box elements can be further classified into distinct subset families according to divergent uracil or cytosine nucleotide content (particularly at the −3 upstream nucleotide) (Fig. 5D; Supplemental Fig. 5B) and divergent conservation profiles (see below). Notably, the distinction between cytosine (CUNAN) and uracil (UUNAN) B-box elements is not limited to the branchpoint motif; rather, each motif family is preferentially associated with distinct types of intron–exon architecture (Fig. 5E,F; Supplemental Fig. 5D–F). Cytosine B-box motifs are associated with cytosine-rich polypyrimidine tracts (PPTs) and high GC% introns and downstream exons. Conversely uracil B-boxes associate with uracil-rich PPTs and low GC% introns and exons. Although uracil B-boxes are characterized by the lowest exonic and intronic GC content, they exhibit the greatest difference in GC content between the exons and their 5′ flanking introns ($P < 0.001$ in all comparisons after Bonferroni correction). In contrast the opposite results are found for the cytosine motif ($P < 0.001$ in all comparisons after Bonferroni correction) (Fig. 5F; Supplemental Fig. 5E,F). These alternate intron–exon architectures share characteristics with those proposed by Amit et al. (2012) to segregate exons according to splicing by either intron- or exon-definition mechanisms.

Competition between splice elements helps decide exon inclusion, with strong splice sites associated with constitutive splicing; whereas competition between weaker splice sites results in alternative splicing (Mullen et al. 1991). Using predicted U2 snRNA base-pairing as a measure of branchpoint strength, we investigated the relationship between branchpoint strength and the splicing status of its associated downstream exon. We confirm that B-box elements exhibiting strongest U2 binding affinity are enriched at constitutively spliced exons (1.26-fold, unpaired *t*-test, *P*-value 1.32×10^{-9}), whereas weaker motifs are associated with alternatively spliced exons (1.67-fold, unpaired *t*-test, *P*-value 3×10^{-4}). Focusing on specific types of alternative splicing events, we find a global shift toward weaker B-box elements for skipped exons ($P < 0.001$), but not within retained introns (one-way ANOVA-Kruskal-Wallis test with Dunn's post-test). B-boxes associated with alternatively spliced exons are also more conserved (1.90-fold, $P = 3.1 \times 10^{-5}$, unpaired *t*-test) than counterparts at constitutive exons, a trend similarly observed for other splicing elements (Sorek 2007).

A model in which B-box strength influences splicing outcome further predicts that when multiple branchpoints can be utilized to splice an exon, branchpoints with stronger U2 binding affinity should out-compete those with weaker B-boxes. Consistent with this hypothesis, we find a positive correlation between B-box strength and branchpoint selection frequency at exons with multiple branchpoints (Spearman r = 0.1709, $P < 0.0001$). Closer examination further suggests that splicing outcome is only impacted once B-box strength crosses a threshold of potential U2 hydrogen bonds, and nonadenosine branchpoints are outcompeted by adenosine branchpoints (one-way ANOVA with Tukey correction for multiple testing) (Fig. 5G).

In addition, we also identified 102 branchpoint sites that conform to the branchpoint sequence UCCUUR$\underline{A}$Y and splice sites of *RNU12*-dependent introns that are spliced by the minor spliceosome (Supplemental Fig. 6A; Turunen et al. 2013).

Conservation of branchpoints and surrounding elements

To investigate branchpoint conservation, we analyzed syntenic nucleotides across 100 vertebrate species (Blanchette et al. 2004).

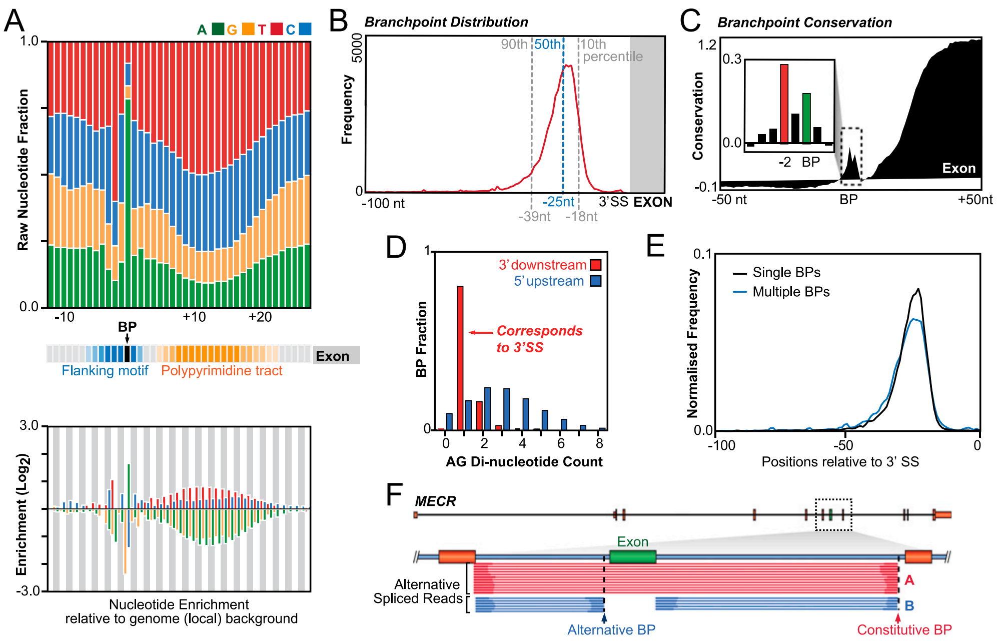

Figure 4. Sequence context and usage of branchpoints. (*A*) Distinct nucleotide composition (*upper* panel) of branchpoints and downstream PPT. Fold enrichment of nucleotide composition relative to local genome background (*lower* panel). (*B*) Frequency distribution of branchpoints relative to the 3′ splice site (dashed lines indicate 10th, 50th, and 90th percentiles). (*C*) Vertebrate conservation profile (100 species) across 100-nt window shows conservation peak around the branchpoint (boxed). Downstream elevated conservation is due to exon sequence. Detailed *inset* shows relative conservation of the branchpoint and flanking −2 upstream nucleotide (red). (*D*) AG dinucleotide counts upstream of branchpoints (blue) and in the downstream intervening region (red) show that the majority of AG dinucleotides in intervening regions correspond to the 3′ splice site. (*E*) Frequency distribution of singleton (black) and multiple (blue) branchpoints relative to the 3′ splice site. (*F*) Alternative splicing of a cassette exon (green) in the *MECR* gene is revealed by lariat reads which have a common 5′ intron termini but utilize branchpoints associated with different 3′SS (red and blue arrows). Blue lariat reads include the cassette exon into the mRNA, while the red lariat reads exclude the cassette exon.

We found that 57.7% of branchpoints were conserved between human and fish lineages, 66.1% within mammal lineages and 95.2% were conserved within primate lineages. B-box elements also exhibit a distinct evolutionary signature, with enrichment for nucleotide substitutions across mammalian lineages that maintain the ability of the B-box to base pair with the U2 snRNA (Fig. 5C).

Among most B-box families, the −2 U nucleotide is the most conserved followed by the branchpoint adenosine (Supplemental Fig. 6B). These nucleotides were especially highly conserved in B-box motifs matching the canonical yeast (*S. cerevisiae*) motif (CUAAC). In contrast, B-boxes without a branchpoint adenosine displayed little or no increase in conservation above the local background. Comparison of the common CUNAN (cytosine) and UUNAN (uracil) motifs revealed similar levels of conservation at the −2 uracil and the branchpoint, but the predominance of uracil nucleotides surrounding the UUNAN B-boxes was also associated with the higher conservation of these nucleotides (Fig. 5E).

We also investigated the relationship between branchpoints and the conservation of associated exons. Similar to the results for exonic GC content, we find a significant difference in exonic nucleotide conservation depending upon the nucleotide composition of the B-box. Exons flanked by CUNAN motifs are significantly less conserved than exons flanked by a UUNAN motif (Mann-Whitney *U*-test, $P < 2.2 \times 10^{-16}$) (Fig. 5E; Supplemental Fig. 6C). The opposite phenomenon is observed for the 3′ splice site strength, where exons flanked by CUNAN show a significantly stronger signal ($P = 8.34 \times 10^{-11}$), suggesting a compensatory mechanism and the interplay between different splicing elements for the accurate splicing of exons (Supplemental Fig. 6D; Dewey et al. 2006).

Human genetic variation at branchpoints

Mutations that abolish branchpoints can result in exon skipping and aberrant splicing with disease consequences (Li et al. 1998; Khan et al. 2004; Padgett 2012). Although branchpoints are refractory (~3.1-fold) to common single nucleotide polymorphisms (SNPs) relative to surrounding sequences (Fig. 6A), we estimate an average ~53 branchpoints are mutated within an individual's genome (The 1000 Genomes Project Consortium 2010).

In contrast to the depletion of common SNPs at branchpoints, we observe a strong enrichment (16.5-fold) for SNPs associated with disease (Fig. 6B; Supplemental Fig. 7A; Forbes et al. 2011; Stenson et al. 2012). In most cases, these disease-associated mutations result in the abolition of conserved adenosine branchpoint nucleotides (Supplemental Table 3). Mutation of the branchpoint can result in exon skipping. For example, we identified a branchpoint nucleotide within the *RB1* gene that when mutated, results in the skipping of the downstream exon in patients with retinoblastoma (Fig. 6C; Houdayer et al. 2008; Zhang et al. 2008). Similarly, we identified a branchpoint within the *MET* oncogene that when deleted, results in skipping of exons encoding the juxtamembrane domain, resulting in *MET* activation in lung adenocarcinoma (Onozato et al. 2009).

Common human genetic variation also exhibits selective constraint in B-box elements, with enrichment for SNPs that maintain base-pairing with the U2 snRNA (Fig. 6D). Notably, the wobble-base pairing ability to pair with two possible nucleotides also protects B-box elements from transition mutations (C-T and A-G). Transition mutations, which are the most common mutational process (65.6% of mutations) in the human genome, do not

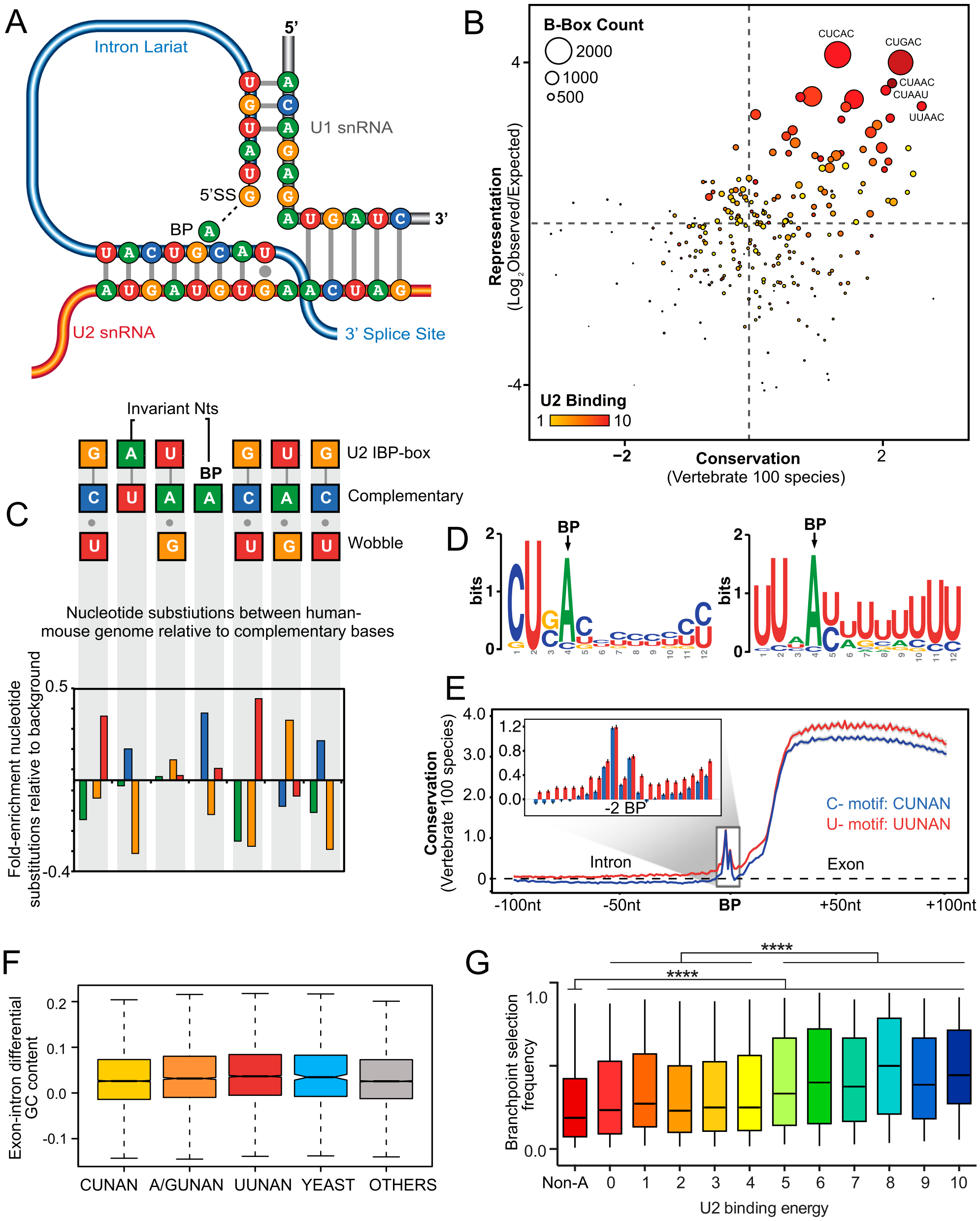

Figure 5. Branchpoint motifs and conservation. (*A*) Schematic illustrating base-pairing of consensus branchpoint flanking sequence to U2 snRNA IBP-box. (*B*) Bubble plot indicating the overrepresentation and conservation of pentamer sequences overlapping branchpoints. Color indicates predicted U2 binding strength and circle radius is proportional to total motif count. Overrepresented motifs, such as CUGAC, exhibit both high conservation and predicted U2 binding strength. (*C*) Enrichment for substitutions that maintain U2:B-box binding. The *upper* schematic shows the predominance of G and U bases within the U2 snRNA IBP box that can bind via complementary or wobble base-pairing to two different possible opposing nucleotides in the B-box. The *lower* histogram indicates the fold enrichment for nucleotide substitutions at syntenic bases within the B-box between the mouse and human genome. Fold enrichment is normalized for background rates of nucleotide substitutions between mouse and human genomes. We observe enrichment for nucleotide substitutions that maintain complementary or wobble base-pairing between the B-box and U2 snRNA. (*D*) Example motifs identified de novo in sequences flanking the branchpoint (branchpoint at +4 nt). (*E*) Average nucleotide conservation score (phyloP 100 vertebrates) for 100 nucleotides flanking the branchpoints. U motifs indicated in red; C-motif indicated in blue. Canonical CUAAC motif is not included in order to examine only derived motifs. Shaded areas represent the 95% confidence intervals. (*Inset*) Twenty nucleotides around the branchpoint, error bars are 95% confidence intervals. (*F*) Box-whisker plot (5%–95% range) of GC% differential between branchpoint introns and associated downstream exons for various families of B-box elements: (Yeast) CUAAC; (others) motifs without a branchpoint adenosine. (*G*) Box-whisker plot (2.5–97.5% range) of relationship between U2 binding energy and branchpoint selection frequency (per exon). (Non A) Nonadenosine branchpoints. Summary of significant differences shown. (****) $P <$ 0.0001; one-way ANOVA with Tukey correction for multiple testing.

compromise the ability of B-box elements to base pair with the U2 snRNA (Fig. 5A; Zhao and Boerwinkle 2002).

In addition, there is a depletion of adenine and guanine polymorphisms (1.2-fold) between the branchpoint and 3′ splice site. This is likely due to the selection against polymorphisms that can form cryptic splice sites that compete with bona fide 3′ splice sites and cause aberrant splicing (Smith et al. 1993). In contrast, 40% (an ∼10.4-fold enrichment over expected) of disease associ-

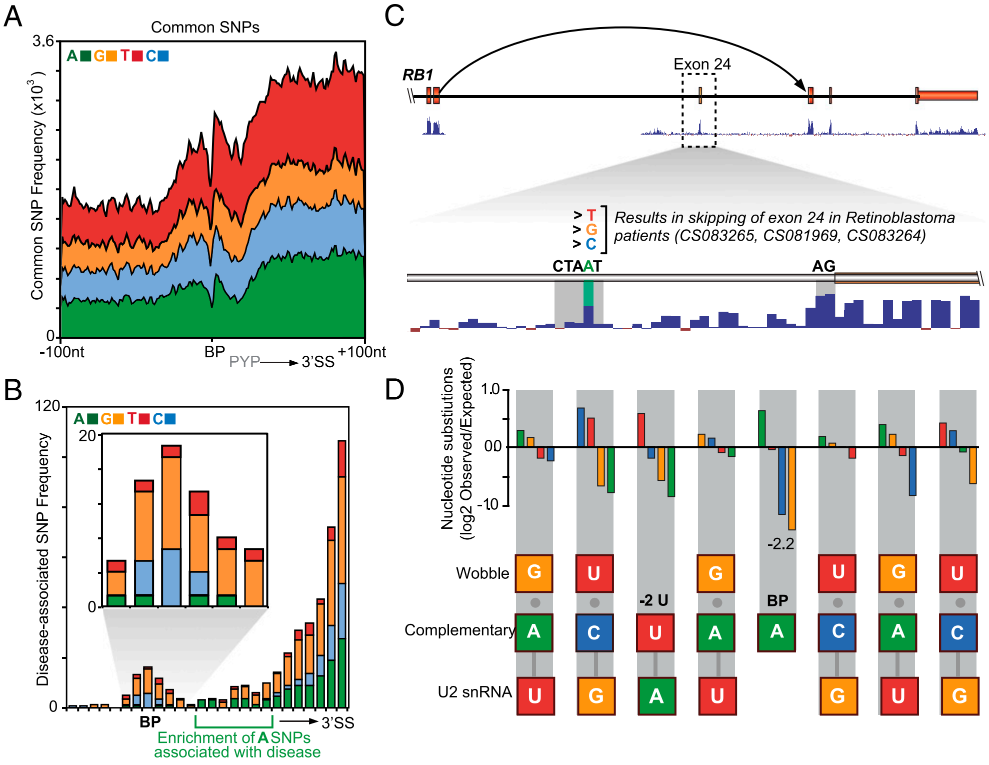

Figure 6. Common and disease-associated genetic variation at branchpoints. (*A*) Frequency distribution of common SNPs across a 200-nt window shows that branchpoints are refractory to common genetic variation (note that downstream exon enrichment is confounded by SNP ascertainment bias). (*B*) Frequency distribution of disease-associated SNPs (with *inset* showing detail) indicates enrichment at branchpoint. (*C*) UCSC Genome Browser view of nucleotide identified as a branchpoint (A; green) in the Retinoblastoma gene (*RB1*). Mutations at this nucleotide have been associated with exon 24 skipping (dashed arrow) in retinoblastoma patients (Houdayer et al. 2008; Zhang et al. 2008). High conservation (Vertebrate 46way) of branchpoint indicated by *lower* blue histogram. (*D*) Histogram indicates fold enrichment for nucleotide substitutions (SNPs) between individual human genomes. Fold enrichment indicates the observed rate of SNPs relative to the expected genome background rate of nucleotide substitution. This demonstrates enrichment for SNPs that maintain complementary or wobble binding between the B-box and U2 snRNA. U2 snRNA and B-box base-pairing nucleotide possibilities are indicated in the *lower* schematic.

ated SNPs occurring downstream from branchpoints result in the formation of a new AG dinucleotide that can potentially act as a cryptic 3′ splice acceptor site (Supplemental Fig. 7B).

Branchpoint prediction and the evolution of branchpoint usage

The identification of explicit B-box motifs allows for the prediction of additional branchpoints in instances in which overrepresented motifs occur in intronic sequences (see Methods). We identified a further 202,646 B-boxes that when combined with our experimentally determined annotations, expands the branchpoint annotations to account for 52.3% of human exons. Predicted B-boxes exhibit sense strand asymmetry and a similarly constrained 3′ proximal peaked distribution as observed for known branchpoints (Fig. 7A). Furthermore, predicted B-boxes exhibit a higher evolutionary conservation (2.1-fold) than surrounding sequence, with only moderately less conservation (~24% lower) than experimentally determined motifs (Fig. 7A; Supplemental Fig. 7D). Similar to known branchpoints, these predicted sites are also refractory to common SNPs and enriched for disease-associated SNPs (Supplemental Fig. 7A,C; Supplemental Table 3). This predictive approach may be necessary to identify branchpoints for lowly expressed exons that are difficult to identify even after sequence capture (Supplemental Fig. 7D) and provide a mechanistic hypothesis for ~1.6% of total disease-associated intronic noncoding DNA variants (Stenson et al. 2012).

The predictive power of B-boxes also extends to other genomes, permitting us to consider usage and innovation of branchpoint motifs in multiple metazoan lineages. Instances of B-boxes were identified in the introns of nine model organisms and assessed according to sense-strand asymmetry and 3′ splice site proximal distribution to enable delineation of lineage-specific trends (Fig. 7B). For example, the most common predicted human CUGAC B-box exhibits an enriched peaked distribution and strand asymmetry within vertebrate lineages, but it is poorly represented within other lineages analyzed. In contrast, the B-box motif, CUAAU is enriched in older human genes that are shared across the vertebrate lineage (1.2-fold, $P = 6 \times 10^{-7}$) (Zhang et al. 2010). This ancient motif exhibits a conserved distribution in all eukaryote lineages analyzed; but while this motif is associated with 26% of *D. melanogastor* exons for which motifs were predicted, its prevalence progressively decreases in the vertebrate lineage, being associated with only 9.1% of human exons (Fig. 7C; Lim and Burge 2001).

The changing usage of branchpoints in different metazoan lineages led us to investigate branchpoint usage in primate specific exons, specifically those associated with *Alu* repeats. In the inverted orientation, the A-rich region between the left and right arms of an *Alu* element and its 3′ polyA tail create cryptic polyT PPTs (Supplemental Fig. 8A). Previous studies have demonstrated the exonization of *Alu* elements requires minimal additional mutations (Lev-Maor et al. 2003; Sorek et al. 2004).

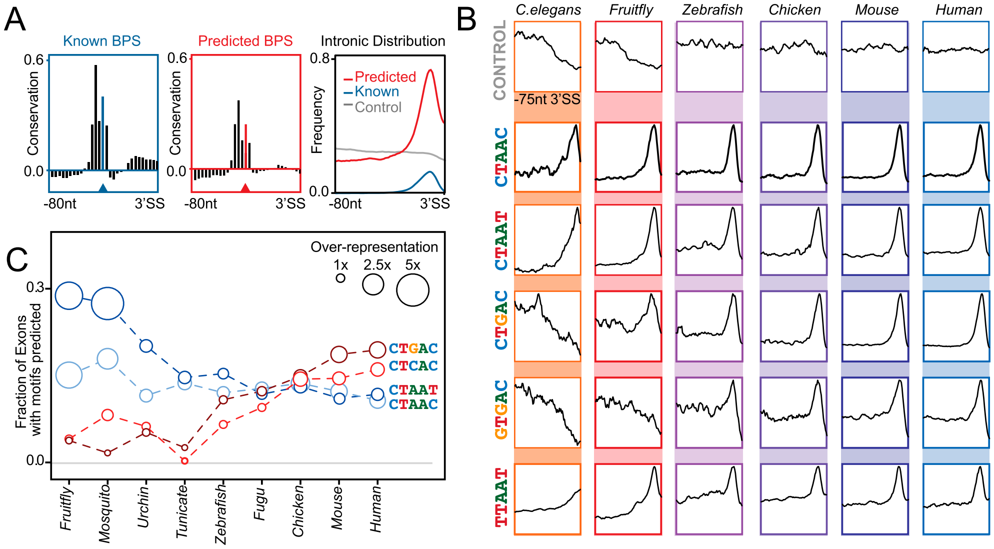

Figure 7. Branchpoint prediction and usage in different species. (*A*) Vertebrate conservation profile (100 species) of known (blue, *left* panel) and predicted (red, *middle* panel) branchpoint sequence (BPS) motifs. Frequency distribution (*right* panel) of known (blue) and predicted (red) motifs relative to the 3′ splice site, with matched scrambled control (gray) indicated. (*B*) Frequency distribution of common branchpoint motifs across a 75-nt window upstream of the 3′ splice site from the range of model organism genomes reveals differential usage of branch motifs across different lineages. (*C*) Bubble plot/histogram indicates the proportion of exons from model organism genomes that are associated with selected motifs. Circle radius is proportional to motif overrepresentation against the expected background in each organism.

Utilizing our experimentally determined branchpoints, we identified 154 (90% of which were adenosine) associated with exonized inverted *Alu*s. Exonized right and left *Alu* arms utilized different sets of B-boxes; 48% of left arm motifs were CUAAU (Supplemental Fig. 8B,C) compared to 1% of right arm B-boxes ($P <$ 0.001, 1000 permutations). The CUAAU B-box is present in the *Alu* inverted consensus sequence just 5′ of the internal polyT tract. CUAAU is an ancient B-box motif with very high U2 binding capability; its presence beside a PPT supports the hypothesis that *Alu* elements are "pre-exons" well placed for inclusion into mature RNA transcripts (Sorek et al. 2004). In contrast, no single position dominates *Alu* right arm exonization, the most common B-box UUUAU (12% of motifs) reflecting instead the T-rich nature of the sequence, with branchpoints present both within the inverted *Alu* polyA tail or close by (Supplemental Fig. 8C,D).

Left arm B-boxes show significantly stronger U2 binding strength ($P < 0.0001$ unpaired *t*-test) (Supplemental Fig. 8E) in contrast to other splicing elements, which have stronger right arm signals (Gal-Mark et al. 2008). Given the impact of B-box strength on splicing outcome (see above), the B-box may compensate for the other left-arm splicing elements and allow exonization.

Discussion

The 5′ and 3′ splice sites of more than ∼300,000 exons have been identified within the human genome to date (Harrow et al. 2012). In contrast, only a few hundred corresponding branchpoints have been collectively described in literature (Taggart et al. 2012; Bitton et al. 2014), and the identification and analysis of branchpoints has been absent from ambitions for a full catalog of functional genetic elements.

The extreme rarity of branchpoint-containing intron lariats has made them almost intractable to genomic analysis. Previous attempts to experimentally determine branchpoints has relied on laborious, low-throughput RT-PCR or 2D-gel purification of lariats (Gao et al. 2008; Awan et al. 2013) or identified several hundred branchpoints from large-scale RNA-sequencing atlases (Taggart et al. 2012; Bitton et al. 2014). In contrast, RNA CaptureSeq and RNase R digestion significantly enrich for intron lariats, allowing genome-wide branchpoint detection in wild-type cells without blocking intron debranching (Bitton et al. 2014). Another advantage of our approach, including the requirement for a diagnostic mismatch at the branchpoint, is the identification of branchpoint nucleotides with both greater accuracy and confidence, allowing us to identify nonstandard branchpoints, including those without an adenosine or distal from their 3′ exon.

This genome-wide annotation provides a unique opportunity to analyze branchpoint features, conservation, and distribution. Although the branchpoint nucleotide itself is strictly constrained in distance from the 3′ splice site and exhibits a strong selection preference for adenosine, the B-box sequences flanking the branchpoint exhibit the full sequence diversity afforded the expanded base-pairing possibilities provided by the keto-nucleotide dense U2 snRNA. These distinct features of B-box elements were previously observed for microRNA seed sequences and may reflect common strategies for sequence specific RNA base pair interactions (Nelson and Green 1989; Wang 2013).

Similar to microRNA selection, the observed heterogeneity of B-box elements may enable the sequence-specific selection of alternate branchpoints by the spliceosome, resulting in splicing regulation and exon inclusion. Although the U2 snRNA branchpoint interacting sequence is strictly conserved throughout eukaryotic lineages (Marz et al. 2008), the density of nucleotide modifications within this branchpoint interacting sequence could further modulate base pair binding possibilities (Yu et al. 1998). Furthermore, it was recently shown that different U2 snRNAs genes exhibit cell-specific expression with mutations to a single U2 snRNA resulting in the disruption of a subset of alternative splicing events, including small introns, in a tissue-specific manner (Jia et al. 2012).

Our results provide some insight into the importance of branchpoint signals and how they are integrated into the wider mechanisms that determine productive splicing. Branchpoints with strong U2 binding (strong B-boxes) outcompete those with weak B-boxes and those without a branchpoint adenosine, to specify exon inclusion (Fig. 5G). In addition, U2 binding strength positively correlates with both B-box occurrence and conservation (Fig. 5B; Supplemental Fig. 5C), supporting the importance of the B-box to efficient splicing. Despite this, splicing is resistant to the impact of B-box mutations. The flexible nature of U2 binding allows transition mutations without a significant impact on B-box strength, whereas our results demonstrate many weaker B-box motifs can participate in branchpoint selection, albeit with decreased efficiency. Furthermore, many genes have multiple branchpoints that can be utilized, providing some redundancy to branchpoint selection.

Our results support previous suggestions that the common −2 U and branchpoint A (UnA) are central to the branchpoint motif (Gao et al. 2008), with these nucleotides showing the highest levels of conservation and motif occurrences as well as depletion of SNPs and overrepresentation in disease. One of the earliest steps in spliceosomal assembly is the binding of SF1 to the branchpoint, a process for which mammalian SF1 requires only the UnA motif, providing a mechanistic explanation for the importance of the −2 U nucleotide (Berglund et al. 1997).

B-box families preferentially associate with distinct classes of intron–exon architecture that can be distinguished by PPT nucleotide content, GC content, and conservation (Fig. 5D–F; Supplemental Figs. 5d–f, 6b–d). It has been proposed these alternative architectures correspond to intron- and exon-defined splicing mechanisms (Amit et al. 2012). Therefore, B-box motifs contribute a further distinction between these two alternative architectures. This integration of multiple splicing features suggests the coevolution of B-box motifs with the surrounding sequence and their integration into the competitive and compensatory mechanisms that regulate splicing.

Uracil-rich PPTs are ancestral with a shift to cytosine enrichment appearing in birds and increasingly in mammals in association with high GC content introns (Amit et al. 2012). U2AF, which binds PPTs cooperatively with the branchpoint binding SF1, has highest affinity for uracil-rich sequences (Coolidge et al. 1997), suggesting that cytosine-rich PPTs, and hence cytosine B-boxes and associated exons, may require additional splicing enhancer sequences to mediate efficient splicing.

Aberrant splicing is responsible for an estimated 15% of human diseases (Singh and Cooper 2012), with exon skipping resulting from the failure of a splicing element, such as mutation to a branchpoint, being the most common cause. We observe disease-associated genetic variation is enriched at both experimentally determined and predicted branchpoints, where it may interfere with splicing. However, branchpoint selection may also have a much broader contribution toward cancer. A range of cancers, in particular hematological malignancies, has been found to harbor mutations to genes encoding the U2 spliceosome responsible for selecting branchpoints (Yoshida et al. 2011). We expect this resource will inform the mechanistic interpretation of disease-associated intronic noncoding variation and aid investigation into how deregulated branchpoint selection constitutes a novel oncogenic pathway.

Methods

Capture array design

Oligonucleotide probes were designed in conjunction with Dr. Ryan Bannen at Roche/NimbleGen using proprietary bioinformatics to optimize array probe sequence and omit repetitive regions. All human genome (hg19) regions from the 100-nt 5′ and 3′ termini of publicly annotated introns (GENCODE v12 comprehensive assembly) were targeted (Harrow et al. 2012). Regions overlapping an annotated exon or a region of high transcription (as determined from publicly available human K562 RNA-seq alignments) (Djebali et al. 2012) were excluded. This final design covered 36.8 Mb and targets both 3′ and 5′ 100-nt termini for 76.4% (206,747) publicly annotated introns or a single terminus for 90.2% (244,125) introns. Additional control probes, including probes targeting all ERCC controls (Baker et al. 2005), were included within the design to assess CaptureSeq performance. The final design was manufactured on a Custom Sequence 2.1M Array (Roche/NimbleGen; Cat #05329841001). Human genome coordinates (hg19) are provided in Supplemental Data 1.

Capture experiment

Capture sequencing was performed similarly to previously described methods (Mercer et al. 2012, 2014) by combining and modifying the NimbleGen SeqCap EZ Library SR User's Guide V3.0 and the NimbleGen Arrays User's Guide: Sequence Capture Array Delivery v3.2. RNA sequencing libraries of ribodepleted total RNA from three K562 biological replicates were created using the TruSeq Stranded mRNA Sample Preparation Kit (Illumina). Precapture and post-capture LMPCR were performed for an average of nine and 17 cycles, respectively. Precapture and post-capture samples (pools of 3 K562 biological replicates) were each sequenced on a single lane of an Illumina HiSeq. See Supplemental Methods for full details.

RNase R treatment

RNase R (Epicentre) digestion was conducted on batches of 100-ng ribodepleted RNA. The standard digestion procedure was 30 units enzyme: 1 µg RNA for 30 min at 37°C. Mock digestion controls lacking RNase R were also performed. See Supplemental Methods for full details.

Validation of RNase R digestion

Digestion of linear RNAs in preference to circular RNAs was confirmed by RT-PCR on untreated Ribo-Zero RNA, RNase R-treated RNA, and RNase R negative mock-treated RNA samples. PCR and Sanger sequencing was used to validate the maintenance of circular multiexonic RNAs within *FBXW4* and *MAN1A2* identified previously (Salzman et al. 2012), while linear RNAs from these same genes were degraded. The fold depletion of linear RNAs was also measured by quantitative real-time PCR (qPCR). Primer sequences are listed in Supplemental Table 4. See Supplemental Methods for full details.

RNase R RNA-seq library preparation

RNA sequencing libraries were made with the TruSeq RNA Sample Preparation v2 Kit (Illumina). Samples were sequenced on an Illumina HiSeq. See Supplemental Methods for full details.

Alignment to identify branchpoint nucleotides

The alignment approach to identify branchpoints is based on the Bowtie 2 read aligner (Langmead and Salzberg 2012) and TopHat2 splice junction mapper (Kim et al. 2013). This pipeline proceeds as follows (illustrated in Supplemental Fig. 3):

Sequenced reads (**.fastq** file) was firstly aligned to the human genome using TopHat2:

```
$ tophat2 –x hg19.index –g GENCODE v12.comprehensive.gtf \
-1 sequences.1.fastq -2 sequences.2.fastq
```

Reads aligning to the reference genome are omitted from further analysis.

An index corresponding to unaligned reads is then assembled (**unaligned_reads.index**). The 5′ (23 nt) sequence of each unique intron (**23nt_5′introns.fa**, using GENCODE v12 comprehensive assembly) is then aligned to the unaligned read index:

```
$ bowtie2 -x unaligned_reads.index -U 23nt_5′intron.fa
```

Unaligned reads with no match to an intron 5′ sequence are omitted. For reads to which a 5′ intron sequence aligns, the sequence downstream to the region aligning to the 5′ intron sequence is trimmed (**trimmed_reads.fa**). The sequence that remains following trimming is required to be longer than 20 nt and is then aligned to the reference human genome:

```
$ bowtie2 -x introns.index -U trimmed_reads.fq
```

The **.sam** output is then analyzed for intronic alignments. Read alignments are required to occur <250 nt of the 3′ splice site of an intron, whose 5′ termini is required to match the original 5′ sequence that was trimmed from the read (i.e., both the splice 5′ intron sequence and trimmed read alignment are required to derive from a single intron). The 3′ nucleotide of the final alignment indicates the predicted branchpoint nucleotide.

As a secondary filter for spurious alignments, we then generated a lariat junction index centered on predicted branchpoints (**lariat.index)**. This lariat junction index comprises the 100 nt upstream of the branchpoint nucleotide followed by the 100-nt sequence from the matched intron 5′ termini, together constituting the expected sequence that traverses the intron lariat junction for each branchpoint. Lariat sequences were required to have <80% homology with the human genome. We then realigned all reads that do not align to the genome:

```
$ bowtie2 -x lariat.index -U unaligned_reads.fq
```

The **.sam** output was filtered for reads requiring a full-length and unique match, with a requisite 20-nt minimum overlap across the branch junction. This provides the final annotation of branchpoints across which lariat reads align.

Identification of sequence errors at branchpoint nucleotides

Reverse transcription across the 2′ to 5′ linkage between the branchpoint and 5′ intron nucleotide is associated with mismatch, insertion, and deletion errors (Vogel et al. 1997).

Mismatch errors were identified within sequenced reads using **samtools calmd** (v1.18) and to determine the **MD/NM** tags that indicate sequence mismatch (Li et al. 2009). Sequence errors were required to correspond to the central branchpoint nucleotide.

Insertion and deletions were identified using:

```
$ bowtie2 -x lariat.index -U unaligned_reads.fq
```

with standard output producing insertion and deletion coordinate (**insertion.bed, deletions.bed**) files. Insertions or deletions were required to occur exactly at the branchpoint nucleotide or, when stranded sequencing was used, be no longer than 3 nt and initiate coincident with the 2′ to 5′ linkage.

Alternative splicing events

Sequenced reads may encompass alternative splicing events. Intron lariats that fully overlapped annotated exons indicate alternative splicing events.

Sequenced reads providing direct evidence for alternative splicing events could be identified as follows: First, reads containing a single unique 5′ intron sequence, joined to multiple unique alignments within 250 nt of the 3′ splice site of the same gene model (using GENCODE v12 transcript ID). Second, reads containing a single match to a branchpoint joined to multiple unique 5′ intron sequences.

Lists of human skipped exons and exons containing retained introns were obtained from http://miso.readthedocs.org/en/fastmiso/annotation.html (Katz et al. 2010). See Supplemental Methods for full details.

Quantification of branchpoint selection

Unique read alignments to the lariat junction index (see “Alignment to identify branchpoint nucleotides” above) provide a raw count of sequence coverage across branchpoint junctions. Alignments were normalized according to combined library size. Analysis was focused on K562 and HeLa.

Branchpoint selection frequency was the read counts for a branchpoint divided by the total number of counts for all branchpoints associated with that same exon. Exons were defined as having dominant branchpoint(s) if the maximum minus the median percentage counts was ≥30%. See Supplemental Methods for full details.

Branchpoint validation by RT-PCR

Nested primer sets (Sigma-Aldrich) were designed to amplify the branchpoint for each chosen candidate after reverse transcription. Purified, amplified DNA was ligated into pGem-T Easy (Promega) and transformed into *E. coli* (Bioline). Sanger sequencing confirmed branchpoints. See Supplemental Methods for full details.

Motif identification

We used the MEME SUITE (Bailey et al. 2009) for de novo motif identification using the following parameters:

```
meme 20nt_sequence_flanking_BPs.fa –dna –minw 5 –o BP_motif
```

To identify genome-wide instances of identified motifs, we used FIMO (Grant et al. 2011) with the default parameters:

```
fimo BP_motif.txt hg19.fa
```

We also identified the fold enrichment of core pentamer sequences flanking branchpoints compared to background (frequency of matched pentamer sequence in a 20-nt window 10 nt directly upstream of the branchpoint). See Supplemental Methods for full details.

Motif strand asymmetry

Strand asymmetry was determined by the frequency of motifs on the sense strand and within 100 nt of the 3′ splice site, relative to motif frequency on combined sense and antisense strand and within 100 nt of the 3′ splice site (i.e., 1 indicates 100% occurrence on sense strand).

Motif intron distribution

To provide a measure of 3′ biased intronic distribution for predicted motifs, the mean frequency of sense motifs within −20 to −50 nt relative to the 3′ splice site was compared to the mean sense motif frequency across the entire intron length.

Nucleotide substitution rate

The nucleotide substitution rate across vertebrate lineages and human genetic variation was determined for sequences flanking branchpoints. The rate of change for each nucleotide flanking branchpoints against reference human sequence was determined using 100-species Vertebrate MULTIZ Alignment (**.maf**) from UCSC (Karolchik et al. 2014). Background nucleotide substitution rate was determined from the sequence upstream (~25 nt) to each branchpoint. Nucleotide substitution rate for human genetic variation relative to reference was determined using dbSNP (build 37, http://www.ncbi.nlm.nih.gov/SNP/), with total nucleotide substitution rate providing background.

Conservation of motifs and surrounding sequence

Human nucleotide conservation score for all branchpoints and associated sequences was retrieved from UCSC Genome Browser (PhyloP Basewise Conservation with 46- or 100-species Vertebrate MULTIZ Alignment) (Blanchette et al. 2004; Pollard et al. 2010). We computed the average nucleotide conservation for 100 nucleotides flanking the branchpoint. Additionally, for each exon with a characterized branchpoint, we computed the average phastCons score across vertebrates (Siepel et al. 2005). Likewise we used MaxEntScan (http://genes.mit.edu/burgelab/maxent/Xmaxentscan_scoreseq_acc.html) (Yeo and Burge 2004) to compare the strength of 3′ splice sites, depending upon B-box composition.

If multiple branch points were identified for an exon, we selected the site with the strongest support.

The phastCons scores of exons as well as the 3′ splice site MaxEntScan scores classified according to the nucleotide composition of their associated B-box motif were compared using a Mann-Whitney *U* test (with Bonferroni correction). See Supplemental Methods for full details.

Motif mapping

Instances of branchpoint pentamer motifs can be identified from genome sequence and gene assemblies. We determined genome coordinates corresponding to instances of motif sequences within the human introns. Motifs coordinates were identified using the findMotif from the Kent source utilities UCSC Tool kit (Karolchik et al. 2014) (http://genomewiki.ucsc.edu/index.php/Kent_source_utilities) that finds exact matches to motif sequence. Identified motifs that overlapped known introns (GENCODE v12 comprehensive assembly) were retained for further analysis. Genome coordinates for motif sequence were also identified in a range of model organism genomes as above. Genome sequences and gene models were downloaded from the UCSC Genome Browser (http://hgdownload.soe.ucsc.edu/downloads.html) as follows: *C. elegans* (ce10; WormBase), *D. melanogastor* (dm3; FlyBase), *D. rerio* (danRer7; RefSeq), *G. gallus* (galGal4, RefSeq), and *M. musculus* (mm10; RefSeq).

U2 binding energy

U2 binding energy measures the number of hydrogen bonds modeled between the motif sequences to the canonical branchpoint binding sequence in the U2 snRNA. We used the Vienna RNA (v2.07) package (Lorenz et al. 2011), RNA duplex script to determine the optimal hybridization structure between U2 snRNA sequence (GUGUAGUA) and the motif (with branchpoint nucleotide removed). Predicted binding energy is determined from the sum of hydrogen bonds forming between complementary motif and U2 snRNA nucleotides.

Gene evolutionary age

The evolutionary age of genes was retrieved from Zhang et al. (2010). Gene names were paired to GENCODE attributes for analysis. Fisher's exact test with multiple hypothesis correction was performed to ascribe significance to enrichments for genes at each lineage.

Branchpoints for exonized *Alu* elements

We downloaded all RepeatMasker *Alu* elements from UCSC (June 17, 2014) and identified all GENCODE exons with an inverted *Alu* element overlapping the 5′ of the exon. We utilized our set of branchpoint-closest 3′ exon associations to identify *Alu* exons with identified branchpoints. See Supplemental Methods for full details.

Branchpoint overlapping disease SNPs

We used the following data sets to determine the overlap between branchpoints and human variation. Common SNPs (dbSNP 137) that are found in >1% of the humans (Sherry et al. 2001), were downloaded from NCBI. Cancer associated SNPs were downloaded from COSMIC (http://cancer.sanger.ac.uk/cancergenome/projects/cosmic/) (Forbes et al. 2011). Mutations and SNPs associated with disease were downloaded from HGMD (Stenson et al. 2012).

Bioinformatics

A number of bioinformatics tool suites were used during analysis. These include BEDTools (Quinlan and Hall 2010), Kent Source Tools, and internal perl/python scripts. Data was downloaded through the UCSC Genome Browser (Karolchik et al. 2014). Statistical analysis and graphing was performed with GraphPad Prism (http://www.graphpad.com/) and R (R Core Team 2013) (http://www.r-project.org/).

Data access

Sequenced RNA-seq libraries have been submitted to the NCBI Gene Expression Omnibus (GEO; http://www.ncbi.nlm.nih.gov/geo/) under accession number GSE53328. Capture probe regions and a list of experimentally determined branchpoints are available as Supplemental Data.

Acknowledgments

The authors would like to thank the following funding sources: Australian National Health and Medical Research Council Fellowships (631668 to J.S.M., also supporting T.R.M. and M.B.C., 631542 to M.E.D., and APP1072662 to M.B.C.); a Queensland State Government (National and International Research Alliance Program, to L.K.N.); and an Australian Research Council Discovery Career Research Award (DE120101916 to R.J.T.). We would also like to thank Roderic Guigo (Centre for Genomic Regulation, Barcelona), IMB Core sequencing facility, Patrick Danoy, Jeffrey Jeddeloh, Ryan Bannen (Roche/NimbleGen), and Daniel Gerhardt (Mater Research) for technical advice and assistance with capture sequencing. We would also like to thank Claudia Wehrspaun for analysis advice and Professor Stephen Rader and Yang Li for manuscript advice. This work was supported by the NeCTAR Genomics Virtual Lab (GVL). The contents of the published material are solely the responsibility of the administering institution, a participating institution, or individual authors and do not reflect the views of NHMRC.

Author contributions: M.B.C. performed RNase R treatments; M.B.C. and J.C. performed sequence capture; S.B.A. and M.B. performed RT-PCR/Sanger validation; T.R.M. performed bioinformatic analyses; M.B.C. and W.H. contributed to bioinformatic analyses; and T.R.M., R.J.T., M.B.C., M.E.D., L.K.N., and J.S.M. conceived the study and prepared the manuscript.

References

The 1000 Genomes Project Consortium. 2010. A map of human genome variation from population-scale sequencing. *Nature* **467:** 1061–1073.

Amit M, Donyo M, Hollander D, Goren A, Kim E, Gelfman S, Lev-Maor G, Burstein D, Schwartz S, Postolsky B, et al. 2012. Differential GC content between exons and introns establishes distinct strategies of splice-site recognition. *Cell Rep* **1:** 543–556.

Awan AR, Manfredo A, Pleiss JA. 2013. Lariat sequencing in a unicellular yeast identifies regulated alternative splicing of exons that are evolutionarily conserved with humans. *Proc Natl Acad Sci* **110:** 12762–12767.

Bailey TL, Boden M, Buske FA, Frith M, Grant CE, Clementi L, Ren J, Li WW, Noble WS. 2009. MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res* **37:** W202–W208.

Baker SC, Bauer SR, Beyer RP, Brenton JD, Bromley B, Burrill J, Causton H, Conley MP, Elespuru R, Fero M, et al. 2005. The External RNA Controls Consortium: a progress report. *Nat Methods* **2:** 731–734.

Berglund JA, Chua K, Abovich N, Reed R, Rosbash M. 1997. The splicing factor BBP interacts specifically with the pre-mRNA branchpoint sequence UACUAAC. *Cell* **89:** 781–787.

Bitton DA, Rallis C, Jeffares DC, Smith GC, Chen YY, Codlin S, Marguerat S, Bahler J. 2014. LaSSO, a strategy for genome-wide mapping of intronic lariats and branch points using RNA-seq. *Genome Res* **24:** 1169–1179.

Blanchette M, Kent WJ, Riemer C, Elnitski L, Smit AF, Roskin KM, Baertsch R, Rosenbloom K, Clawson H, Green ED, et al. 2004. Aligning multiple genomic sequences with the threaded blockset aligner. *Genome Res* **14:** 708–715.

Coolidge CJ, Seely RJ, Patton JG. 1997. Functional analysis of the polypyrimidine tract in pre-mRNA splicing. *Nucleic Acids Res* **25:** 888–896.

Cornish-Bowden A. 1985. Nomenclature for incompletely specified bases in nucleic acid sequences: recommendations 1984. *Nucleic Acids Res* **13:** 3021–3030.

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

Dewey CN, Rogozin IB, Koonin EV. 2006. Compensatory relationship between splice sites and exonic splicing signals depending on the length of vertebrate introns. *BMC Genomics* **7:** 311.

Djebali S, Davis CA, Merkel A, Dobin A, Lassmann T, Mortazavi A, Tanzer A, Lagarde J, Lin W, Schlesinger F, et al. 2012. Landscape of transcription in human cells. *Nature* **489:** 101–108.

Forbes SA, Bindal N, Bamford S, Cole C, Kok CY, Beare D, Jia M, Shepherd R, Leung K, Menzies A, et al. 2011. COSMIC: mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer. *Nucleic Acids Res* **39:** D945–D950.

Gal-Mark N, Schwartz S, Ast G. 2008. Alternative splicing of *Alu* exons—two arms are better than one. *Nucleic Acids Res* **36:** 2012–2023.

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

Gerstein MB, Bruce C, Rozowsky JS, Zheng D, Du J, Korbel JO, Emanuelsson O, Zhang ZD, Weissman S, Snyder M. 2007. What is a gene, post-ENCODE? History and updated definition. *Genome Res* **17:** 669–681.

Grant CE, Bailey TL, Noble WS. 2011. FIMO: scanning for occurrences of a given motif. *Bioinformatics* **27:** 1017–1018.

Harrow J, Frankish A, Gonzalez JM, Tapanari E, Diekhans M, Kokocinski F, Aken BL, Barrell D, Zadissa A, Searle S, et al. 2012. GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res* **22:** 1760–1774.

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.

Houdayer C, Dehainault C, Mattler C, Michaux D, Caux-Moncoutier V, Pagès-Berhouet S, d'Enghien CD, Laugé A, Castera L, Gauthier-Villars M, et al. 2008. Evaluation of in silico splice tools for decision-making in molecular diagnosis. *Hum Mutat* **29:** 975–982.

Jia Y, Mu JC, Ackerman SL. 2012. Mutation of a U2 snRNA gene causes global disruption of alternative splicing and neurodegeneration. *Cell* **148:** 296–308.

Kapranov P, Drenkow J, Cheng J, Long J, Helt G, Dike S, Gingeras TR. 2005. Examples of the complex architecture of the human transcriptome revealed by RACE and high-density tiling arrays. *Genome Res* **15:** 987–997.

Karolchik D, Barber GP, Casper J, Clawson H, Cline MS, Diekhans M, Dreszer TR, Fujita PA, Guruvadoo L, Haeussler M, et al. 2014. The UCSC Genome Browser database: 2014 update. *Nucleic Acids Res* **42:** D764–D770.

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.

Khan SG, Metin A, Gozukara E, Inui H, Shahlavi T, Muniz-Medina V, Baker CC, Ueda T, Aiken JR, Schneider TD, et al. 2004. Two essential splice lariat branchpoint sequences in one intron in a xeroderma pigmentosum DNA repair gene: mutations result in reduced *XPC* mRNA levels that correlate with cancer risk. *Hum Mol Genet* **13:** 343–352.

Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL. 2013. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol* **14:** R36.

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

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.

Lev-Maor G, Sorek R, Shomron N, Ast G. 2003. The birth of an alternatively spliced exon: 3′ splice-site selection in *Alu* exons. *Science* **300:** 1288–1291.

Li M, Kuivenhoven JA, Ayyobi AF, Pritchard PH. 1998. T→G or T→A mutation introduced in the branchpoint consensus sequence of intron 4 of lecithin:cholesterol acyltransferase (LCAT) gene: intron retention causing LCAT deficiency. *Biochim Biophys Acta* **1391:** 256–264.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25:** 2078–2079.

Lim LP, Burge CB. 2001. A computational analysis of sequence features involved in recognition of short introns. *Proc Natl Acad Sci* **98:** 11193–11198.

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.

Marz M, Kirsten T, Stadler PF. 2008. Evolution of spliceosomal snRNA genes in metazoan animals. *J Mol Evol* **67:** 594–607.

Mercer TR, Gerhardt DJ, Dinger ME, Crawford J, Trapnell C, Jeddeloh JA, Mattick JS, Rinn JL. 2012. Targeted RNA sequencing reveals the deep complexity of the human transcriptome. *Nat Biotechnol* **30:** 99–104.

Mercer TR, Clark MB, Crawford J, Brunck ME, Gerhardt DJ, Taft RJ, Nielsen LK, Dinger ME, Mattick JS. 2014. Targeted sequencing for gene discovery and quantification using RNA CaptureSeq. *Nat Protoc* **9:** 989–1009.

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.

Nelson KK, Green MR. 1989. Mammalian U2 snRNP has a sequence-specific RNA-binding activity. *Genes Dev* **3:** 1562–1571.

Neph S, Vierstra J, Stergachis AB, Reynolds AP, Haugen E, Vernot B, Thurman RE, John S, Sandstrom R, Johnson AK, et al. 2012. An expansive human regulatory lexicon encoded in transcription factor footprints. *Nature* **489:** 83–90.

Onozato R, Kosaka T, Kuwano H, Sekido Y, Yatabe Y, Mitsudomi T. 2009. Activation of MET by gene amplification or by splice mutations deleting the juxtamembrane domain in primary resected lung cancers. *J Thorac Oncol* **4:** 5–11.

Padgett RA. 2012. New connections between splicing and human disease. *Trends Genet* **28:** 147–154.

Pollard KS, Hubisz MJ, Rosenbloom KR, Siepel A. 2010. Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Res* **20:** 110–121.

Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26:** 841–842.

R Core Team. 2013. *R: a language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/.

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

Salzman J, Gawad C, Wang PL, Lacayo N, Brown PO. 2012. Circular RNAs are the predominant transcript isoform from hundreds of human genes in diverse cell types. *PLoS ONE* **7:** e30733.

Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, Sirotkin K. 2001. dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res* **29:** 308–311.

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.

Singh RK, Cooper TA. 2012. Pre-mRNA splicing in disease and therapeutics. *Trends Mol Med* **18:** 472–482.

Smith CW, Porro EB, Patton JG, Nadal-Ginard B. 1989. Scanning from an independently specified branch point defines the 3′ splice site of mammalian introns. *Nature* **342:** 243–247.

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.

Sorek R. 2007. The birth of new exons: mechanisms and evolutionary consequences. *RNA* **13:** 1603–1608.

Sorek R, Lev-Maor G, Reznik M, Dagan T, Belinky F, Graur D, Ast G. 2004. Minimal conditions for exonization of intronic sequences: 5′ splice site formation in *alu* exons. *Mol Cell* **14:** 221–231.

Stenson PD, Ball EV, Mort M, Phillips AD, Shaw K, Cooper DN. 2012. The Human Gene Mutation Database (HGMD) and its exploitation in the fields of personalized genomics and molecular evolution. *Curr Protoc Bioinformatics* **39:** 1.13.11–11.13.20.

Suzuki H, Zuo Y, Wang J, Zhang MQ, Malhotra A, Mayeda A. 2006. Characterization of RNase R-digested cellular RNA source that consists of lariat and circular RNAs from pre-mRNA splicing. *Nucleic Acids Res* **34:** e63.

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.

Turunen JJ, Niemela EH, Verma B, Frilander MJ. 2013. The significant other: splicing by the minor spliceosome. *Wiley Interdiscip Rev RNA* **4:** 61–76.

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

Wang B. 2013. Base composition characteristics of mammalian miRNAs. *J Nucleic Acids* **2013:** 951570.

Will CL, Lührmann R. 2011. Spliceosome structure and function. *Cold Spring Harb Perspect Biol* **3:** a003707.

Yeo G, Burge CB. 2004. Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. *J Comput Biol* **11:** 377–394.

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.

Yu YT, Shu MD, Steitz JA. 1998. Modifications of U2 snRNA are required for snRNP assembly and pre-mRNA splicing. *EMBO J* **17:** 5783–5795.

Zhang K, Nowak I, Rushlow D, Gallie BL, Lohmann DR. 2008. Patterns of missplicing caused by *RB1* gene mutations in patients with retinoblastoma and association with phenotypic expression. *Hum Mutat* **29:** 475–484.

Zhang YE, Vibranovski MD, Landback P, Marais GA, Long M. 2010. Chromosomal redistribution of male-biased genes in mammalian evolution with two bursts of gene gain on the X chromosome. *PLoS Biol* **8:** e1000494.

Zhang Y, Zhang XO, Chen T, Xiang JF, Yin QF, Xing YH, Zhu S, Yang L, Chen LL. 2013. Circular intronic long noncoding RNAs. *Mol Cell* **51:** 792–806.

Zhao Z, Boerwinkle E. 2002. Neighboring-nucleotide effects on single nucleotide polymorphisms: a study of 2.6 million polymorphisms across the human genome. *Genome Res* **12:** 1679–1686.

Received August 12, 2014; accepted in revised form November 20, 2014.

Genome-wide discovery of human splicing branchpoints

Tim R. Mercer, Michael B. Clark, Stacey B. Andersen, et al.

Genome Res. 2015 25: 290-303 originally published online January 5, 2015
Access the most recent version at doi:10.1101/gr.182899.114