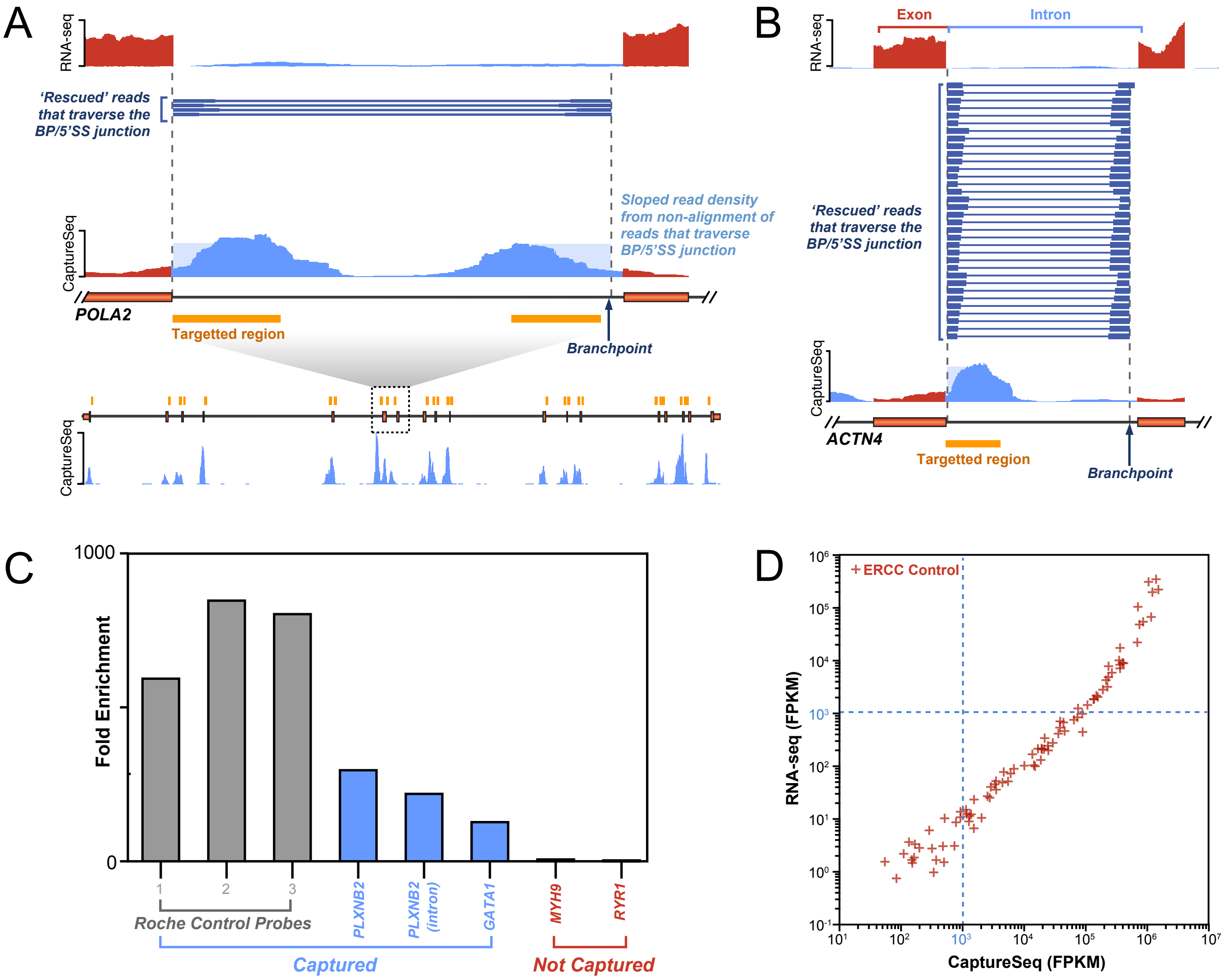
Genome-wide discovery of human splicing branchpoints

Tim R. Mercer, Michael B. Clark, Stacey B. Andersen, Marion E. Brunck, Wilfried Haerty, Joanna Crawford, Ryan J. Taft, Lars K. Nielsen, Marcel E. Dinger and John S. Mattick

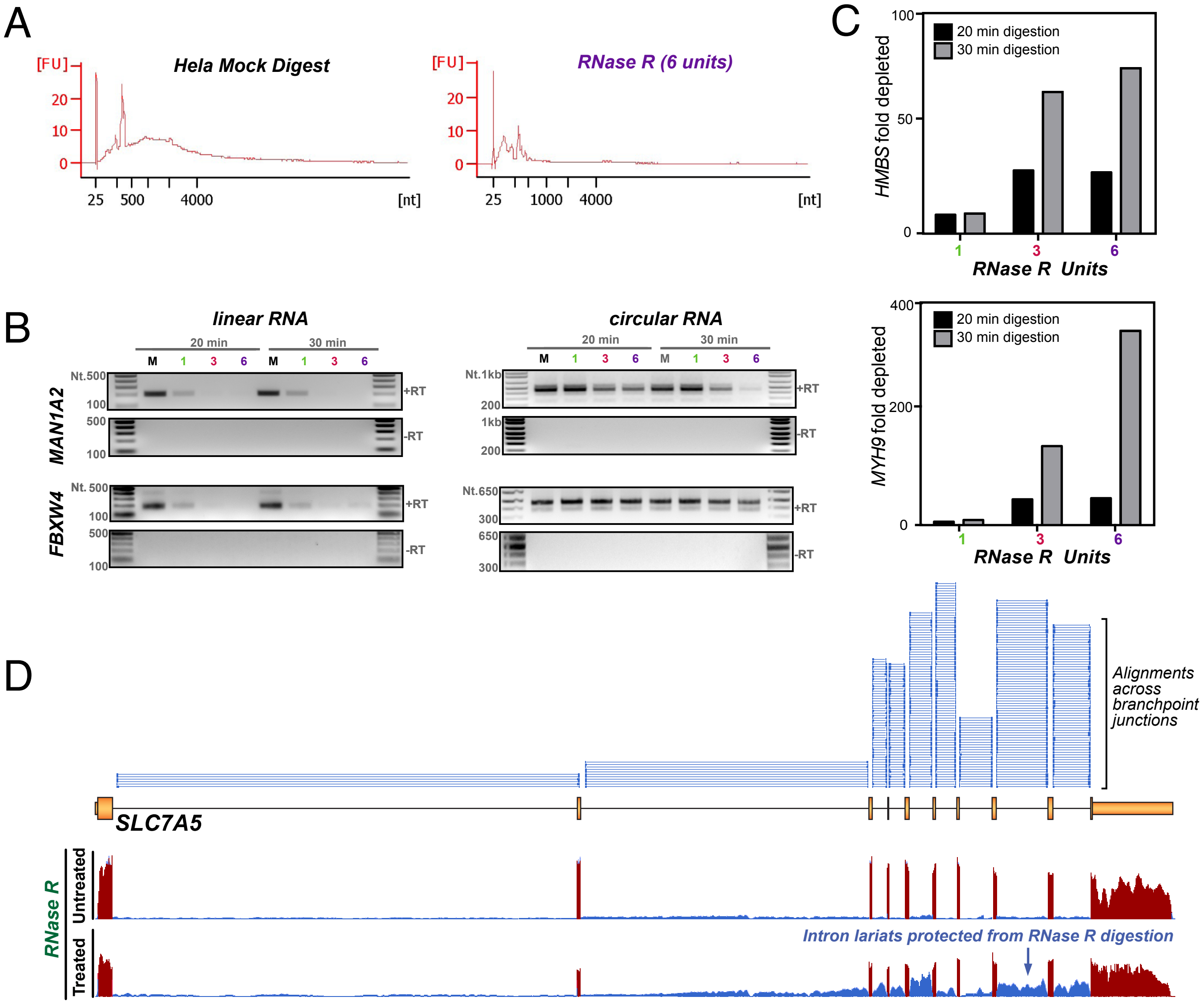
SUPPLEMENTARY INFORMATION

Supplementary Figure 1



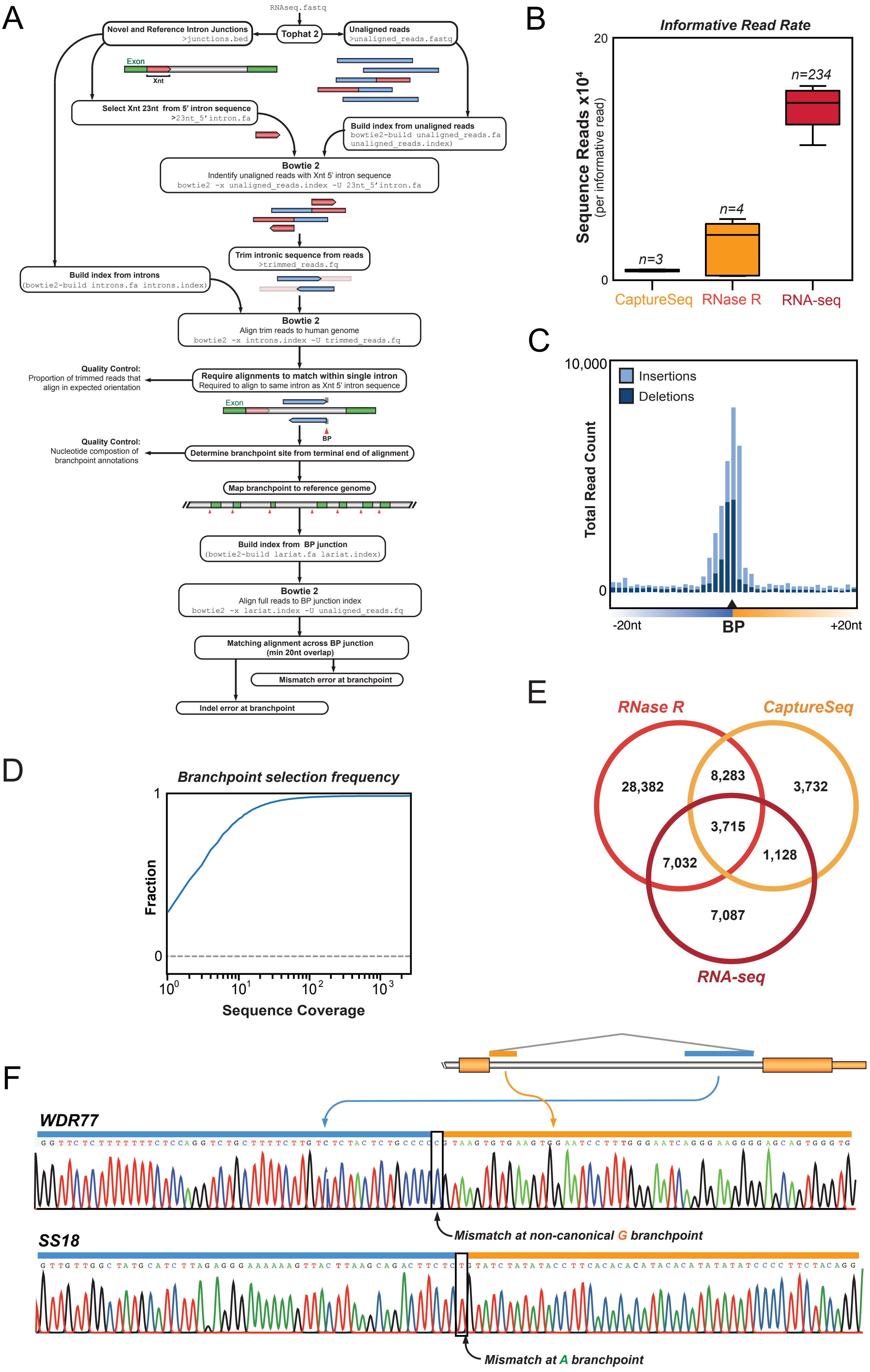
Supplementary Figure 1. CaptureSeq enrichment of intron lariat reads. (a) Probe design targets terminal intronic regions (orange boxes) of *POLA2* gene, resulting in the enrichment of intronic reads (blue) by CaptureSeq (lower histogram) relative to conventional RNA-seq (upper histogram). Read density exhibits a downwards-graded slope as reads that align to the genome (light blue) approach either the 5' splice site or branchpoint as a result of the inability to align reads traversing the branchpoint with conventional alignment. A non-conventional split and inverted alignment strategy 'rescues' reads (dark blue) that align across the 5' splice site and branch-point junction. **(b)** Probes solely targeting the 5' sequence of an *ACTN4* intron is sufficient to enrich reads traversing the lariat junction and determine branchpoint location. **(c)** Quantitative RT-PCR shows fold-enrichment following CaptureSeq for targeted Roche controls (grey; 1-3), *PLXNB2* and *GATA1* genes (blue) and depletion of non-targeted *MYH9* and *RYR1* genes (red). **(d)** Scatter plot of ERCC probe expression (FPKM) shows CaptureSeq exhibits >100 fold higher sampling (indicated by blue dashed line) of ERCC probes relative to conventional RNA-seq.

Supplementary Figure 2



Supplementary Figure 2. RNase R protection of intron lariats. (a) Bioanalyzer (Agilent 2100) traces of ribodepleted HeLa RNA after mock-digested (**left panel**) and following RNase R digestion (6-unit of enzyme per 100ng of ribodepleted RNA; **right panel**). **(b)** Digestion of *MAN1A2* and *FBXW4* linear RNA (**left panel**) in HeLa cells, shown by RT-PCR of *MAN1A2* exons 11-13, or *FBXW4* exons 8-9. Protection of *MAN1A2* and *FBXW4* circular RNA (**right panel**), determined by outwards facing primers in exon 2 (minutes indicate duration of digestion; M: mock digested). **(c)** Fold depletion of *HMBS* and *MYH9* linear transcripts in RNase R treated HeLa sample by qPCR compared to mock digestion. **(d)** Genome browser view showing read alignments from RNase R-digested (lower histogram) and mock-digested (upper histogram) libraries at *SLC7A5* gene loci. RNase R digestion shows depletion of exonic sequences (red) and corresponding enrichment for intron lariats (blue reads). Read alignments (blue) from RNase R libraries that traverse lariat junction indicated above.

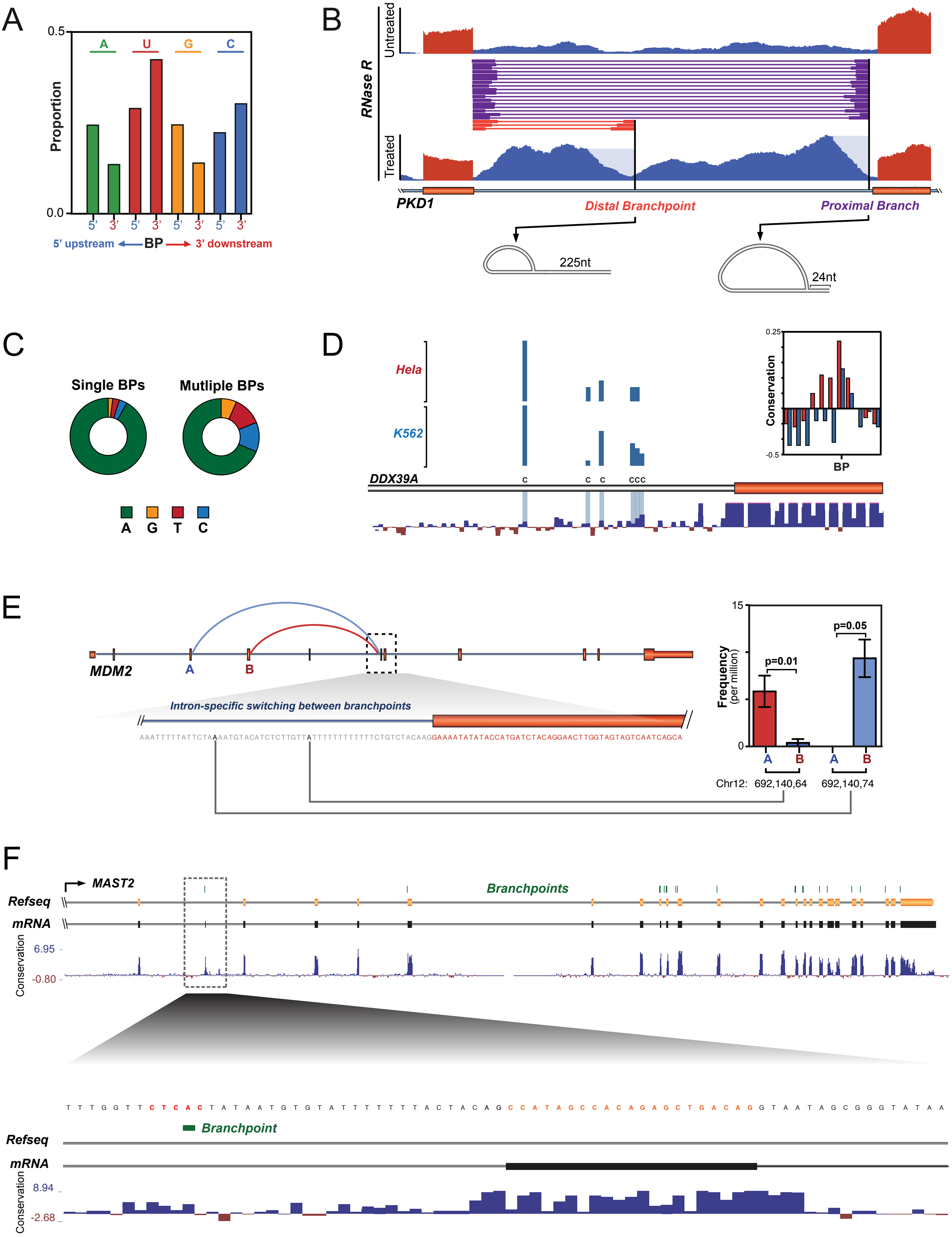
Supplementary Figure 3



```
>junctions.bed
>unaligned_reads.fastq
>23nt_5'intron.fa
bowtie2-build unaligned_reads.fa
unaligned_reads.index)
bowtie2 -x unaligned_reads.index -U 23nt_5'intron.fa
>trimmed_reads.fq
(bowtie2-build introns.fa introns.index)
bowtie2 -x introns.index -U trimmed_reads.fq
(bowtie2-build lariat.fa lariat.index)
bowtie2 -x lariat.index -U unaligned_reads.fq
```

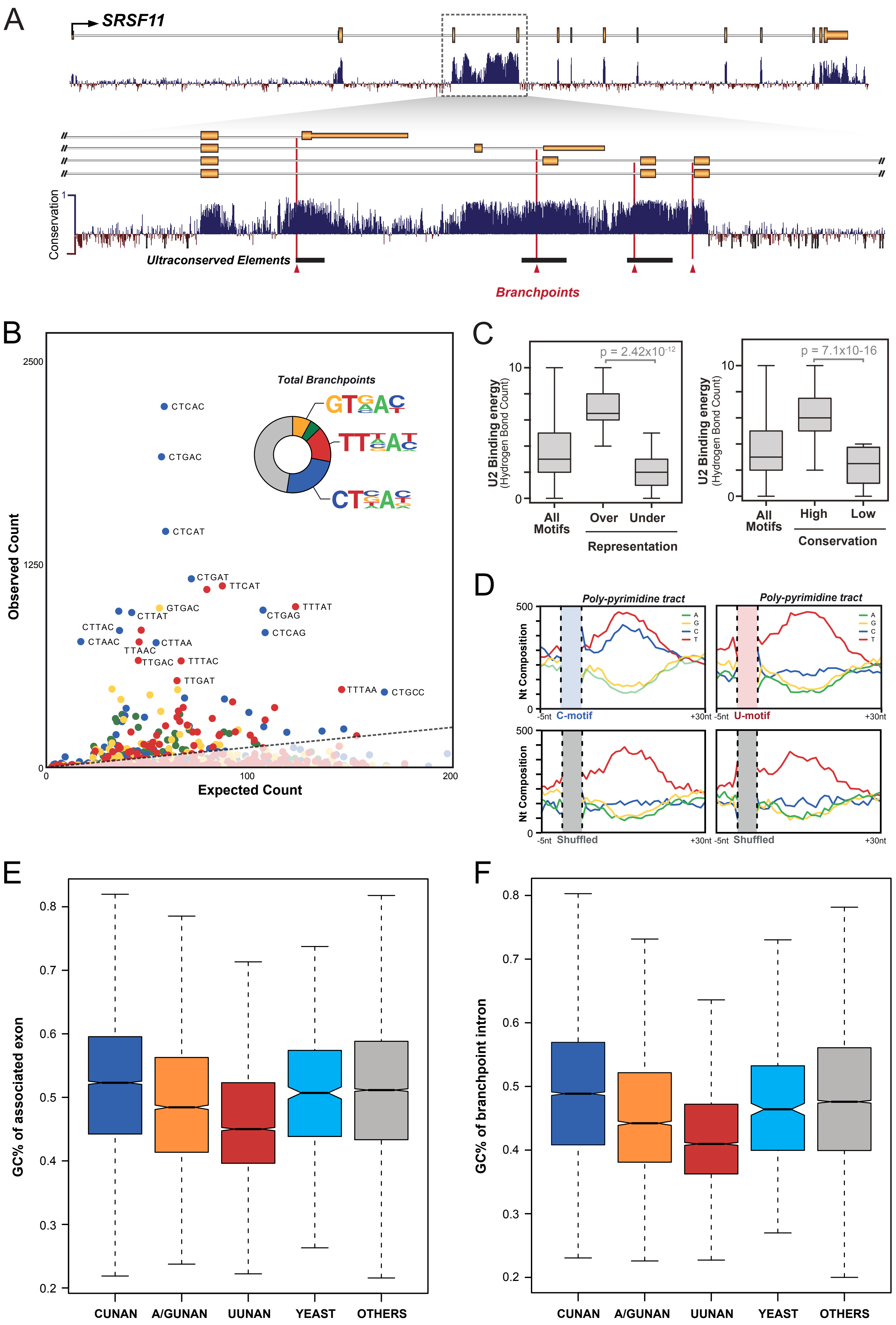
Supplementary Figure 3. Strategy for identification of branchpoints from sequenced reads. (a) Alignment strategy used to identify branch nucleotides. Schematic diagram indicates the approach, tools and parameters employed within the alignment pipeline to identify branchpoints (Further detail is provided within **Methods**). **(b)** Rate of informative reads that traverse a branchpoint per total reads sequenced for CaptureSeq, RNAse R and conventional RNA sequencing. **(c)** Histogram indicates rate of insertion/deletion errors at branchpoint nucleotide in sequenced reads that align across lariat junction (center point defined from overlapping sequence reads with single mismatch error). **(d)** Cumulative frequency of sequence read coverage over branchpoint annotations shows dynamic quantitative range. **(e)** Venn diagram indicating the relative contribution and overlap between each approach to identify high-confidence branchpoints. The number of branchpoints annotated using each approach was CaptureSeq (16,858 – 3 libraries, one cell type), ENCODE (18,962 – 234 libraries, many cell types) and RNAse R (47,412 – 4 libraries, two cell types). **(f)** Validation of branchpoint locations by RT-PCR and Sanger sequencing. Examples genes were selected for targeted intron-specific amplification across branchpoint junction, followed by Sanger sequencing of amplicon. Sequenced amplicons are split at the branchpoint (blue and orange bars) and inverted for alignment. Examples of mismatch nucleotide incorporation (boxed) shown in chromatograph at canonical adenosine branchpoint nucleotide in *SS18* **(lower chromatograph)** and mismatch nucleotide incorporation at non-canonical guanine branchpoint nucleotide in *WDR77* **(upper chromatograph)**.

Supplementary Figure 4



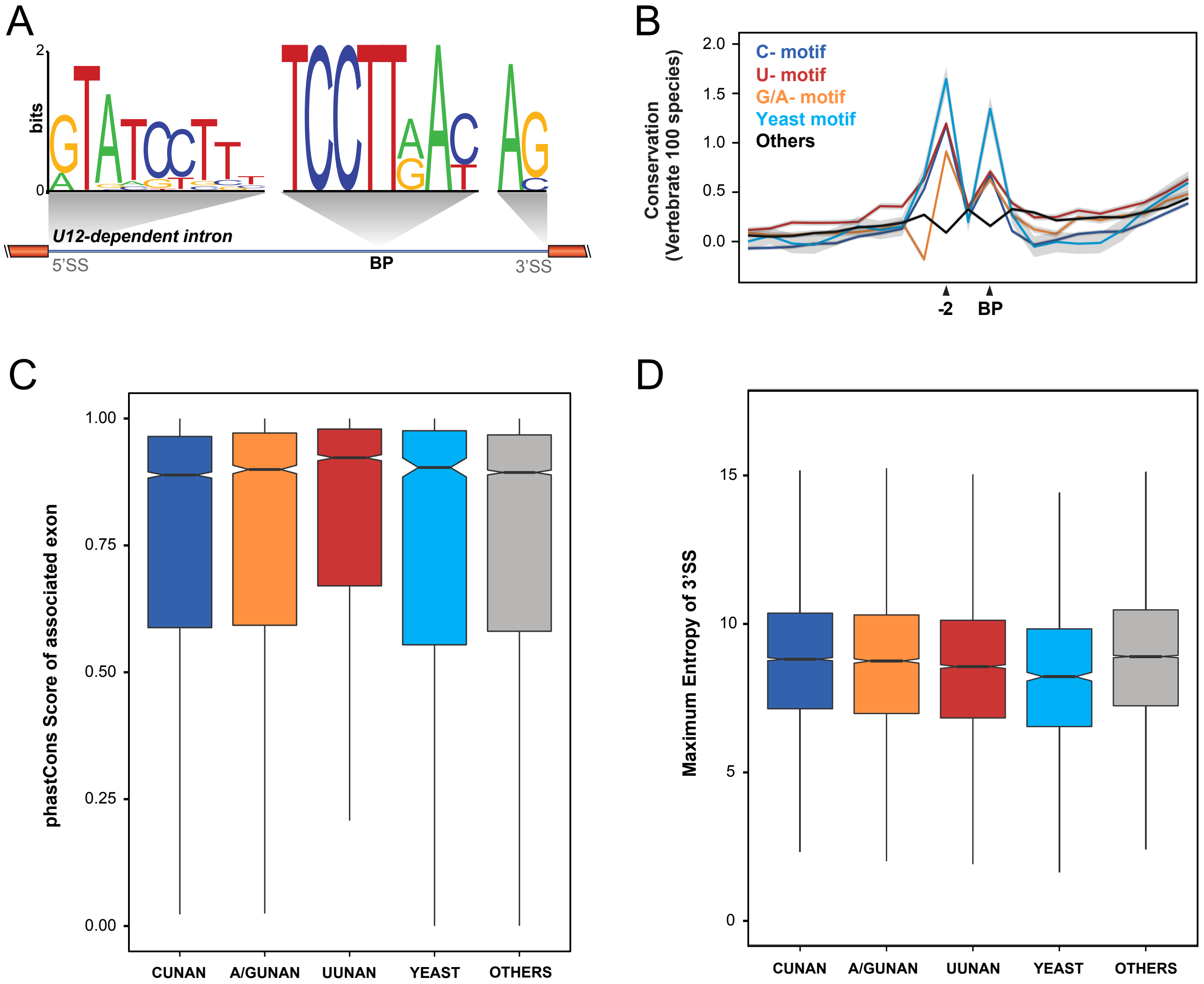
Supplementary Figure 4. Branchpoint features. (a) Relative proportion of nucleotides within matched upstream (5'; blue) and intervening downstream regions (3'; red) relative to branchpoint shows depletion of purines (A,G) following branchpoint. **(b)** Unusual example of two distal branchpoints employed within single *PKD1* intron. Use of the proximal branchpoint forms a conventional lariat structure with 24 nucleotide downstream tail, while use of the distal branchpoint forms a lariat with a 225 nucleotide downstream trailing sequence. Reads traversing distal (red) and proximal (blue) branchpoints are indicated, and both lariat structures are supported by the characteristic graded slope in the RNase R distribution 5' to of the branchpoints. **(c)** Pie chart indicates nucleotide composition of singleton (left) and multiple (right) branchpoints, showing lower adenine selection preference for multiple branchpoints. **(d)** Genome browser view of cytosine branchpoint cluster in *DDX39A* intron. **(Upper panel)** histogram shows quantitative selection of cytosine branchpoints in K562 and HeLa cell-types. **(Inset)** Inset indicates conservation (vertebrate 100-way) of multiple cytosine branchpoints (blue) relative to total branchpoints (red). **(e)** Quantitative profiling illustrates distinct branchpoint selection preferences in alternative splicing of *MDM2* gene. We observe a significant difference ($p<0.05$, paired t-test, n=4, error bars SD) in branchpoint selection between two alternative upstream intronic 5' termini. **(f)** Confirmation of inclusion of a micro-exon in *MAST2* gene. Exon missing from Refseq and GENCODE v19 gene catalogues but supported by mRNA evidence. Branchpoint nucleotide in green. Branchpoint pentamer B-box motif in red, exonic sequence in orange.

Supplementary Figure 5



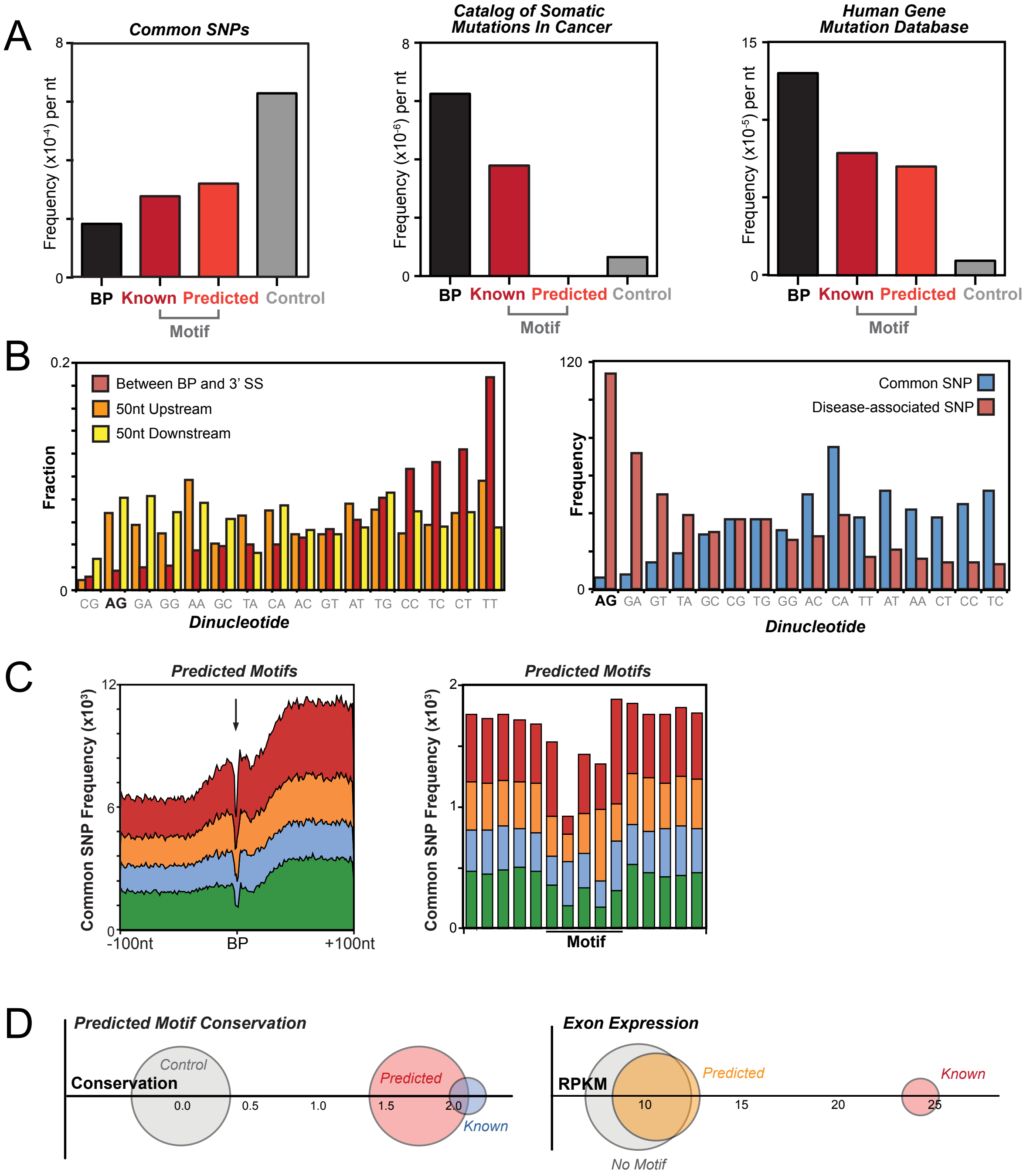
Supplementary Figure 5. Sequence context of branchpoints. (a) Genome browser view of *SRSF11* gene, showing the overlap of branchpoints (red arrows) with ultraconserved elements associated with auto-regulatory non-productive alternative splicing of SR proteins (Lareau *et al*. 2007). **(b)** Scatter-plot indicates the observed relative to expected frequency of pentamer sequences overlapping branchpoints. Points colored according first nucleotide and dashed line indicates no enrichment. **(Inset)** Pie chart indicates the relative proportion of branch sites corresponding to common CUnAn, UUnAn and GUnAn branch motifs. **(c)** Box-whisker plot (min-max range) showing predicted U2 binding energy is enriched for over-represented **(upper panel)** or highly conserved **(lower panel)** branchpoint motifs relative to all motifs (unpaired t-test). **(d)** Relative cytosine and thymine proportions of polypyrimidine tracts downstream to *de novo* identified U- and C-motif branchpoints (**upper panels**). Shuffled control motifs shown below (**lower panel**). **(e,f)** Box plots of GC% of downstream exons **(e)** and the branchpoint containing introns **(f)** for various families of B-box elements. Plots displays 5-95% range. Yeast: CUAAC canonical motif, invariant in *S.cerevisiae*; Others: B-box motifs without a branchpoint adenosine.

Supplementary Figure 6



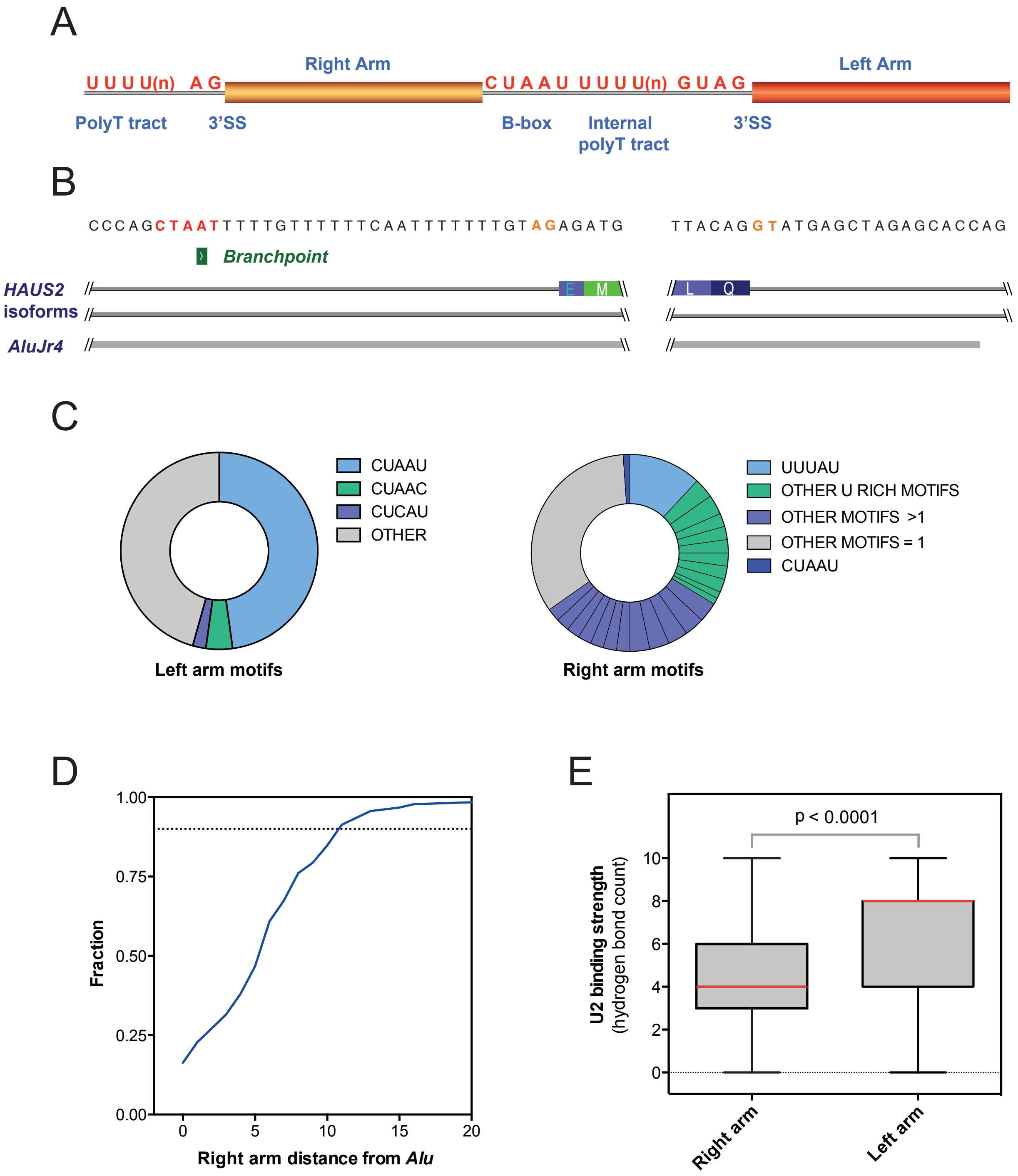
Supplementary Figure 6. Branchpoint and associated sequence conservation. (a) Nucleotide motifs of splicing elements (3' and 5' splice sites and branchpoints) identified within U12-dependent introns. **(b)** Average nucleotide conservation score (phylOP 100 vertebrates) for 20 nucleotides centered on the B-box motif. U motif: UUNAN; C-motif: CUNAN; G/A-motif: G/AUNAN; Yeast: CUAAC canonical motif, invariant in *S.cerevisiae*; Others: B-box motifs without branchpoint adenosine. Shaded areas are 95% confidence intervals. **(c,d)** Box plots of exon conservation **(c)** and 3'SS strength **(d)** of exons associated with various classes of B-box elements. Plot displays Min to Max range.

Supplementary Figure 7



Supplementary Figure 7. Branchpoint association with disease. (a) Overlap frequency of common single nucleotide polymorphisms (SNPs; **left panel**) and disease associated SNPs (Catalog of Somatic Mutations in Cancer, **middle panel**; Human Gene Mutation Database, **right panel**) with branchpoints, known and predicted overlapping sequence motifs and matched scrambled controls. **(b; left panel)** Dinucleotide frequency within region between the branchpoint and 3' splice site indicates depletion of AG. **(right panel)** Frequency of di-nucleotides generated by disease (red) and common (blue) SNPs within intervening regions between branch-point and 3' splice site. The formation of an AG dinucleotide that may act as a cryptic 3' splice site are associated with disease. **(c)** Frequency distribution of commons SNPs in relation to predicted branchpoint motifs with detailed inset shown **(middle panel)**. **(d; left panel)** Average conservation of known (blue) and predicted (red) motif sequences with matched scrambled control (grey) indicated. Circle radius is proportional total motif count in human introns. **(right panel)** Average expression of exons preceded by known, predicted or no motif.

Supplementary Figure 8



Supplementary Figure 8. *Alu* element exonization. (a) Schematic of inverted *Alu* element consensus sequence. In inverted orientation, the 3' polyA tract and internal polyA sequence form cryptic polypyrimidine tracts. Inverted *Alu* elements contain one or more AG motifs that can form a 3' splice site (SS). CUAAU B-box also shown. **(b)** Example of exonized left arm *Alu* with CUAAU branchpoint motif (red) in *HAUS2* gene. 3'SS and 5'SS nucleotides labeled in orange. **(c; left panel)** Branchpoint motifs utilized for *Alu* left arm exonization. CUAAU is in *Alu* consensus sequence, CUAAC and CUCAU use occurred at diverged CUAAU sites. No other motif was utilized more than once. **(right panel)** Branchpoint motifs utilized for *Alu* right arm exonization. UUUAU was most frequently utilized motif. Other U rich motifs (minimum 3/5 U nucleotides) displayed in green, other motifs utilized more than once or once in purple and grey respectively. **(d)** Distance between branchpoints and *Alu* element for right arm exonizations. 90% of branchpoints are less than 11 nucleotides 5' of *Alu* element, while 16% are within the *Alu* element sequence (commonly from an A within the polyU tract). **(e)** Box-whisker plot (Tukey) showing predicted U2 binding energy is higher for *Alu* element left arm B-box motifs (unpaired t-test). Red bars show median values.

SUPPLEMENTARY METHODS

Cell culture and RNA extraction

Human K562 and HeLa cells were cultured in RPMI and DMEM respectively plus 10% fetal bovine serum and penicillin/streptomycin at 37°C, 5% $C0_2$. K562 cells were collected by centrifugation, washed with PBS, centrifuged again and lysed in TRIzol (Life Technologies). HeLa cells were lysed in TRIzol by scraping of the cell culture flask. A standard TRIzol extraction was conducted and RNA resuspended in 200ul of RNase free water. Purity was confirmed by NanoDrop (Thermo scientific). RNA was DNase treated with Turbo™ DNase (Life Technologies), repurified with phenol/chloroform, nanodroppped again and the integrity of the RNA confirmed by Agilent 2100 Bioanalyzer (Agilent Technologies). The DNA free nature of the RNA was confirmed by the absence of product from a PCR for a section of *NEAT1* gDNA with 200ng of purified RNA for 35 cycles. Five microgram lots of high quality RNA were treated with Ribo-Zero™ (Epicentre) to remove rRNA, purified by RNeasy MinElute (Qiagen) with successful ribodepletion confirmed by Bioanalyzer (Agilent 2100) Pico chip.

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 100nt 5' and 3' termini of publicly annotated introns (GENCODE v12 comprehensive assembly) were targeted (Harrow et al. 2012). Any 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)) was excluded. This final design covered 36.8Mb and targets both 3' and 5' 100nt termini for 76.4% (206,747) publicly annotated introns, or a single terminus for 90.2% (244,125) introns. Additional control probes were included within the design to assess CaptureSeq performance. Probes targeting all ERCC controls (Baker et al. 2005) as well as 30 complex spliced genes for internal control of splice junction and transcript assembly were included. Final design, including controls, were manufactured on a Custom Sequence 2.M Array (Roche/NimbleGen. Cat #05329841001). Human genome coordinates (hg19) are provided in **Supplementary Data 1**.

Capture Experiment

Capture sequencing was performed similar to previously described (Mercer et al. 2012; Mercer et al. 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). Different index adaptors (4,6,12) were added to each K562 biological replicate. Library input consisted of ribodepleted RNA from 5 µg original total RNA after quality checking. ERCC RNA Spike-In Control mix 1 (Invitrogen) was added to ribodepleted RNA to give a final ERCC concentration of 1% (replicate 1 and 2) and 1.08% (replicate 3). Library preparation was begun at the fragmentation step by the addition of 9 µl EPH buffer and the standard protocol followed until "Enrich DNA Fragments". The 20 µl obtained at the end of the "Ligate Adaptors" step was increased to 21 µl with resuspension buffer and mixed well. One microliter of this solution was added to 19 µl resuspension buffer in a new 0.3ml PCR plate and mixed. This new plate was utilized for the "Enrich DNA Fragments" step to create amplified test libraries to guide PreCapture LMPCR with the remaining 20 µl.

Precapture LMPCR and QIAquick PCR Purification Kit (Qiagen) were performed as described by the NimbleGen SeqCap EZ Library SR User's Guide V3.0. To test LMPCR yields K562 biological replicate 3 libraries were amplified for 8, 9 or 10 cycles, quantified on the Bioanalyzer (Agilent 2100) and then pooled. K562 biological replicate 1 and 2 libraries were thus amplified for 9 cycles and the yield quantified by Bioanalyzer (Agilent 2100).

Capture hybridization was modified from the NimbleGen Arrays User's Guide: Sequence Capture Array Delivery V3.2. The NimbleGen Hybridization System was set to 42°C and allowed 3 h to equilibrate. Equal nanograms of library from each K562 biological replicate were pooled, some of pooled library was allocated for pre-capture sequencing and 1ug of the pooled library utilized for capture. The library for capture was mixed with 300 µg human Cot-1 DNA (Invitrogen) and 3.34 µl 100 µm TS-INV-HE (hybridization enhancing) Index Oligos (IDT) to each index adaptor plus 1 µl 1000 µm TS-HE Universal Oligo 1 (IDT) in a 1.5 ml tube. Lid of tube was pierced with an 18 gauge needle and the sample dried at 60 °C in a vacuum concentrator. Once dry the tube was given a new lid to prevent contamination through the needle hole. Library/Cot/Oligo mix was resuspended in 11.2 µl of nuclease and nucleic acid free water, vortexed for 10 s and centrifuged at max speed for 10 s. Sample was solubilized at 70 °C for 10 m before repeating vortexing and centrifugation. Nimblegen 2 x SC Hybridization Buffer (18.5 µl) and SC Hybridization Component A (7.3 µl) were added and vortexing and centrifugation repeated. Sample was denatured at 95°C for 10 m. During this time the “Prepare Mixers” procedure from the NimbleGen Arrays User’s Guide was performed with the omission of the compressed gas step. After denaturation samples was vortexed for 10 s, centrifuged at max speed for 10 s and incubated at 42°C until ready to load onto the NimbleGen Hybridization system. The ‘Load and Hybridize Samples’ procedure from the NimbleGen Arrays User’s Guide was performed to begin array hybridization. Hybridization was conducted for ~3 days.

The NimbleGen Arrays User’s Guide protocol was followed to prepare the elution chamber, disassemble, wash and elute the captured DNA from the microarray, with the following modifications. Gasket and elution chamber was setup in a DNA-free laminar flowhood. All washes were conducted with 50 ml buffer in 50 ml falcon tubes. MinElute purification (Qiagen) was conducted using a microfuge. Captured DNA was recovered with two 30 µl elutions from a MinElute column and the final volume adjusted to 60 µl.

Post-capture LMPCR and cleanup was performed similar to the NimbleGen SeqCap EZ Library SR User’s Guide V3.0, with the following modifications. LMPCR was run using 5x Phusion buffer for 17 cycles. Each PCR contained 70 µl of master mix and 30 µl of captured DNA. Quantity and quality of amplified captured DNA was determined by Bioanalyzer (Agilent 2100).

Enrichment of captured transcripts was measured by qPCR using SYBR Green PCR Master Mix and real time cyclers (Applied Biosystems). Enrichment was determined by the ratio of transcript abundance between the pre and post capture samples using equal nanograms of each. Enrichment was determined for three NimbleGen capture controls and 3 sequences captured specifically by the branchpoint array (primer sequences are listed in **Supplementary Table 4**). Two transcripts not targeted by the capture array were also tested to examine the specificity of capture. Concentrations and volumes were as per the NimbleGen SeqCap EZ Library SR User’s Guide V3.0.

Precapture and post capture samples (3 K562 biological replicates) were each sequenced on a single lane of an Illumnia® HiSeq.

RNase R treatment

RNase R (Epicentre) digestion was conducted on batches of 100 ng ribodepleted RNA. A number of digestion conditions were tested with the standard digestion procedure being 30U enzyme: 1 µg RNA for 30 min at 37°C. Mock digestion controls lacking RNase R were also performed. RNA was purified by RNeasy MinElute and digestion of RNA by RNase R confirmed by Bioanalyzer (Agilent 2100) Pico chip.

Validation of RNase R digestion

Digestion of linear RNAs in preference to circular RNAs was confirmed by RT-PCR. Reverse transcription utilized the SuperScript™ III cDNA synthesis kit (Life Technologies) using random hexamers and equivalent proportions of input RNA. Reverse transcription was conducted on untreated Ribo-Zero RNA, RNase R treated RNA and the RNase R negative mock-treated RNA sample. PCR was used to validate the maintenance of circular multi-exonic RNAs within *FBXW4* and *MAN1A2* identified previously (Salzman et al. 2012) (using outwards facing primers in a single exon), while linear RNAs from these same genes were degraded. Sanger sequencing was performed to validate the identity of the multi-exonic *FBXW4* and *MAN1A2* circular RNAs. The fold depletion of linear RNAs was measured by quantitative real-time PCR (qPCR) against *HMBS* and *MYH9*. qPCR was performed using SYBR Green PCR Master Mix and real time cyclers (Applied Biosystems). Primer sequences are listed in **Supplementary Table 4**.

RNase R RNA-seq library preparation

RNA sequencing libraries were made with the TruSeq® RNA Sample Preparation v2 Kit (Illumina). Given the RNA was previously ribodepleted, library preparation was begun at the fragmentation step and the standard procedure followed. Successful generation of sequencing libraries from digested and RNase R negative mock-digestion samples were confirmed by Agilent 2100 Bioanalyzer. Samples were sequenced on an Illumina® HiSeq. Given the very small amount of RNA remaining after RNase R digestion, multiple digested samples were combined to reach the minimum yield requirement for sequencing library preparation. Where necessary samples digested with different conditions that all provided satisfactory digestion were also pooled.

Alignment to identify branchpoint nucleotide

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 **Supplementary 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’ (23nt) sequence of each unique intron (**23nt_5’introns.fa**, using GENCODE v12 comprehensive assembly) is then aligned to 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 downstream sequences 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 20nt 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 (ie. both the splice 5’ intron sequence and trimmed read alignment are required derive from 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 branchpoint (**lariat.index)**. This lariat junction index comprises the 100nt upstream to the branchpoint nucleotide followed by the 100nt sequence from the matched intron 5’ termini, together constituting the expected sequence to traverses the intron lariat junction for each branchpoint. Lariat sequences were required to have less than 80% homology to human genome. We then re-aligned all reads that do not align to 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 20nt minimum overlap across branch junction. This provides the final annotation of branchpoints across which lariat reads align.

Stranded library preparation (using TruSeq® Stranded mRNA Sample Preparation Kit) was performed for RNase R and CaptureSeq libraries. To provide an indication of false positive alignment rate for each library, we determined number of sequenced reads incorrectly aligning in antisense direction across the branch junction, divided by the read count correctly aligning in the antisense direction. The mean rate across all libraries was reported as the false positive rate for read alignments.

Identification of sequence errors at branchpoint nucleotides

Reverse transcription across the 2'5 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** (v 1.18) 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 from Tophat2:

```
$ 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 exact at the branchpoint nucleotide or, when stranded sequencing was used, be no longer than 3nt and initiate coincident with the 2' to 5' linkage.

Alternative splicing events

Sequenced reads may encompass alternative splicing events. We firstly determined full lariat coordinates, with the matched intron 5' termini indicating start of coordinates and branchpoint nucleotide as stop coordinate. 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; Firstly, reads containing a single unique 5’ intron sequence, joined to multiple unique alignments within 250nt of 3’ splice site of the same gene model (using GENCODE v12 transcript Id); Secondly, reads containing single match to branchpoint, joined to multiple unique 5’ intron sequence.

Lists of human skipped exons and exons containing retained introns were obtained from the annotations of human genome (hg19) alternative events v2.0 as part of the documentation for MISO (http://miso.readthedocs.org/en/fastmiso/annotation.html) (Katz et al. 2010).

Quantification of branchpoint selection

The sequence coverage across lariat sequence can provide a quantitative measure of branchpoint selection. Unique read alignments to the lariat junction index (see Alignment to Identify Branchpoint Nucleotide above) provide a raw count of sequence coverage across branchpoint junctions. Unique read alignments were normalized according to combined library size. Analysis was focused on K562 and HeLa cell-types that afforded deepest coverage.

Analysis of differences in branchpoint selection between cell types was restricted to cases where multiple branchpoints are clustered at single exon. Statistical difference was ascribed using unpaired t-test, with n = 4 individual libraries per cell type and performed using R.

Branchpoint selection frequency was calculated to identify dominant branchpoints and examine the impact of B-box strength on branchpoint selection. Exons associated with multiple branchpoints were filtered to retain those where the branchpoint with maximum use had >3 counts. 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%. The global relationship between B-box strength and branchpoint selection frequency was determined by Spearman correlation. The distributions of branchpoint selection frequencies at each level of B-box strength were also compared by one-way ANOVA with Tukey correction for multiple testing. Other measures and requiring higher maximum counts gave comparable results to those reported.

Branchpoint validation by RT-PCR

Nested primer sets (Sigma-Aldrich) were designed to amplify the branchpoint for each chosen candidate. First-strand cDNA synthesis was performed with 500 ng DNase-treated RNA, using SuperScript II reverse transcriptase (Life Technologies) and outer reverse primer. The first round of PCR was set up using cDNA and Phusion Hot Start Flex DNA polymerase (New England Biolabs) with an outer primer set, and divided into multiple reactions performed at different annealing temperatures. PCR products were pooled, and DNA purified using ISOLATE II PCR and Gel Kit (Bioline). Purified DNA was used as template for a second round of PCR using an inner primer set, divided across different annealing temperatures. PCR products were combined and run on an agarose gel. Bands of interest were excised, DNA extracted using ISOLATE II PCR and Gel Kit (Bioline), and products A-tailed using *Taq* polymerase with Thermopol buffer (New England Biolabs). DNA was ligated into pGem-T Easy cloning vector (Promega), transformed into α-select chemically-competent *E. coli* (Bioline) and incubated overnight on selective plates. Four colonies were harvested for each branchpoint and grown up overnight in selective media. Plasmids were purified using ISOLATE II Plasmid Mini Kit (Bioline) and digested to check for an appropriate insert. Those showing expected digest pattern were Sanger sequenced using a T7 sequencing primer by the Australian Genome Research Facility.

RNA binding proteins

Occupancy coordinates for HNRNP (A1, A2B1, F, H1, M, and U) proteins were retrieved from Huelga et al. (Huelga et al. 2012). The distribution of occupancy sites was determined across 200 nt genome regions centered on the branchpoint nucleotide.

Motif Identification

We employed 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 employed FIMO (Grant et al. 2011) with the default parameters:

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

We also identified the overrepresentation of core pentamer sequences flanking branchpoints that correspond to the U2 snRNA IBP-box element. The observed frequency of pentamer sequences overlapping all unique branchpoint annotations (branchpoint at nucleotide 4) was determined as follows. The expected background frequency was estimated by the frequency of matched pentamer sequence in a 20 nt window 10 nt directly upstream to the branchpoint. The over-representation of each motif sequence was the determined by the fold-enrichment of observed to expected pentamer frequency.

Motif strand asymmetry

Strand asymmetry measures the strand bias of identified branchpoint motifs. 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

Branchpoints exhibit a peaked intronic distribution in close relation to the 3' splice site. 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 alignments (MULTIZ alignment .maf file) downloaded from the UCSC Genome Browser (Karolchik et al. 2014). Background nucleotide substitution rate was determined for sequence upstream (~25nt) matched at 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 Vertebrate MULTIZ Alignment (Blanchette et al. 2004; Pollard et al. 2010). This evaluates individual nucleotides for both accelerated (faster than expected under neutral drift) and conserved (slower than expected evolution). Conservation scores represent -log p-values under a null hypothesis of neutral evolution. To compute the conservation of different branchpoint motifs and surrounding sequence, we computed the average nucleotide conservation for 100 nucleotides flanking the branchpoint. Additionally, for each exon with a characterized branchpoint we computed the average nucleotide phastCons conservation 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 B-box motifs were classified into different families depending upon their nucleotide composition (U motif: UUNA̲N; C-motif: CUNA̲N; G/A-motif: G/AUNA̲N; Yeast: CUAA̲C canonical motif, invariant in *S.cerevisiae*; Others: B-box motifs without a branchpoint adenosine.).

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).

GC% of surrounding sequence and GC% differential

Previous results associated splicing with G+C content (Amit et al. 2012). We compared the nucleotide composition of exons and introns classified according to their branch point motif. For each exon and its 5' flanking intron we calculated the G+C content as well as their difference in GC content. If multiple branch points were identified within an intron, the branch point with the strongest support was selected for analysis.

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), *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. Hydrogen bonds form between G:C(3), A:U(2) and G:U(1) with the branchpoint nucleotide bulging out and being omitted from the pairings. We employed 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 the determined from 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) whom inferred the origination of genes based on the presence or absence of the gene in vertebrate phylogeny (12 lineages). Gene names were paired to GENCODE Attributes for analysis. A fisher–exact test with multiple hypothesis correction was performed to ascribe significance to enrichments for genes at each lineage.

Branchpoints supporting exons not present in gene catalogues

Branchpoints have a restricted distribution of distances from 3’ exons, therefore branchpoints that are distant from gene catalogue exons but which are a standard distance from a non-gene catalogue mRNA-supported exon may represent the splicing of this exon instead. All branchpoints were associated with the closest 3’ exon from GENCODE v19 and GENCODE v19 plus mRNA exons from UCSC (downloaded 18 June 2014), to provide a list of mRNA exons closer in distance to branchpoints than any GENCODE exon. This list was filtered to retain those mRNAs exons more than 20 nt closer to the branchpoint than the GENCODE exon but more than 18 nt from the branchpoint themselves. Next, branchpoint-mRNA pairs were required to utilize a canonical splice acceptor and fit a spliceosomal scanning model of 3’ splice site recognition. Remaining pairs were examined manually to remove any of dubious quality.

Branchpoints for exonized *Alu* elements

We downloaded all repeat masker *Alu* elements from UCSC (17 June 2014) and identified all GENCODE exons with an inverted *Alu* element overlapping the 5’ of the exon. We then utilized our set of branchpoint – closest 3’ exon associations to identify which *Alu* exons had identified branchpoints. Exons were split into left arm exonizations (branchpoint in internal region) and right arm exonizations (branchpoint in polyA(U) tail or outside of *Alu* element) and manually inspected.

Disease SNPs with branchpoints

We employed the following datasets to determine overlap between branchpoints and human variation. Coordinates for common SNPs (dbSNP 137) that are found in greater than 1% of the humans (Sherry et al. 2001), we downloaded from NCBI. Cancer associated SNPs were downloaded from the Catalogue Of Somatic Mutations In Cancer database (http://cancer.sanger.ac.uk/cancergenome/projects/cosmic/) (Forbes et al. 2011). The data is curated from literature, CGP laboratories at the Sanger Institute, UK, TCGA, ICGC and IARC p53 data portal. Mutations and SNPs associated with disease were download from the public Human Gene Mutation Database (HGMD) that show the genomic coordinates of disease-associated variants in the public version of the database (Stenson et al. 2012).

Bioinformatics

A number of bioinformatics tool suites were employed 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/).

SUPPLEMENTARY TABLE AND DATA LEGENDS

Supplementary Table 1. Human genome (hg19) coordinates of branchpoint nucleotides with support from **(i)** exact read match to predicted only, **(ii)** matches and insertion or deletions coincident with branchpoints, **(iii)** matches and sequencing error at branchpoint, **(iv)** branchpoint from initial split and inverted alignment.

Supplementary Table 2. Human genome (hg19) coordinates of branchpoint nucleotides supporting mRNA exons not present in Refseq or GENCODE gene catalogues.

Supplementary Table 3. List of disease-associated SNPs from Human Gene Mutation Database (HGMD) and Catalog of Somatic Mutations in Cancer (COSMIC) that overlap branchpoints, known or predicted motifs.

Supplementary Table 4. List of primer sequences utilized in the study.

Supplementary Data 1. Human genome (hg19) coordinates for capture array probe design (.bed file)**.**

Supplementary Data 2. Human genome (hg19) coordinates for high confidence branchpoint annotations (.bed file)**.**

Supplementary Data 3. Human genome (hg19) coordinates for introns from GENCODE annotations (v12 basic) classified according to single or multiple BP status (.txt file). Columns 1-3 and 5: intron co-ordinates and strand; column 4: GENCODE transcript ID and intron number; column 6: number of branchpoints annotated within the intron; column 7: branchpoint co-ordinates and nucleotide.

SUPPLEMENTARY REFERENCES

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.

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**: W202-208.

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.

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.

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**: D945-950.

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.

Huelga SC, Vu AQ, Arnold JD, Liang TY, Liu PP, Yan BY, Donohue JP, Shiue L, Hoon S, Brenner S et al. 2012. Integrative genome-wide analysis reveals cooperative regulation of alternative splicing by hnRNP proteins. *Cell Rep* **1**: 167-178.

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**: D764-770.

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.

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.

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.

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

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.

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.

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.

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.

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.

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

Yeo G, Burge CB. 2004. Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. *Journal of computational biology : a journal of computational molecular cell biology* **11**: 377-394.

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**.