“Large-scale mapping of branchpoints in human pre-mRNA transcripts *in vivo*.” Taggart et al.

Supplementary Material

1) Supplementary Table 1 (66 disease causing mutations in branchpoints)

2) Supplementary Table 2 (Sequencing results of lariat validation)

3) Supplementary Figure 1 (Evidence for redundant and for non-redundant branchpoints in human pre-mRNA.)

4) Supplementary Figure 2 (Recovery rate of lariats by intron length)

5) Supplementary Figure 3 (Mutational profile of RT at branchpoints)

6) Supplementary Methods:

A) Computational Method
I. Lariat Discovery
II. Analysis
III. Statistical Tests

B) Experimental Methods
I. RT-PCR validation of Lariats

Further resources can be found at http://fairbrother.biomed.brown.edu/data/Lariat/ or fairbrother@brown.edu

Supplementary Table 1

	HGMD ID	Gene Symbol	Original Base	New Base	Sequence
Hypoglycaemia, persistent hyperinsulinaemic	CS961698	ABCC8	A	G	TGGCCTC**A**CTTGTGC
Hypoglycaemia, persistent hyperinsulinaemic	CS050778	ABCC8	G	A	CCTGGGC**G**GTGGGAC
Adenosine deaminase deficiency	CS930753	ADA	G	A	GTTCTCT**G**GTTCCAT
Spherocytosis	CS961473	ANK1	C	T	CTCTCCC**C**GGCCGGC
Hypertriglyceridaemia	CS075066	APOC2	G	A	GCCCCAC**G**GGCTCTC
Primary microcephaly	CS091963	ASPM	A	G	GAATATA**A**TATCTGG
Bardet-Biedl syndrome ?	CS032059	BBS2	A	G	ACTTTTA**A**ATTTGTG
Breast cancer ?	CS045210	BRCA1	T	C	TAACTAG**T**GTTTCTT
Breast cancer	CS982093	BRCA2	A	G	AATTTAT**A**AAGCAGC
Agammaglobulinaemia	CS961496	BTK	A	G	GAGTCTC**A**CTGGTCT
Muscular dystrophy, limb girdle	CS053449	CAPN3	C	G	GCTCTCT**C**TCTTCTT
Cystic fibrosis	CS001829	CFTR	T	C	ACCAACA**T**GTTTTCT
Cystic fibrosis	CS086376	CFTR	G	A	TTGCAAT**G**TTTTCTA
Alport syndrome	CS982128	COL4A4	A	G	GCCTTCA**A**TTTTTTT
Ehlers-Danlos syndrome II	CS982129	COL5A1	T	G	GAGTGAC**T**GACCAGC
Epidermolysis bullosa dystrophica	CS094363	COL7A1	A	G	TGCTCTG**A**TTTCTTC
Cystinosis, nephropathic	CS102107	CTNS	T	C	TCAGCAG**T**AATTAGA
Iron overload	CS042809	CYBRD1	G	C	TCATCCT**G**TTTGTAA
D-2-hydroxyglutaric aciduria	CS050424	D2HGDH	A	G	AAACATG**A**AATTACC
Muscular dystrophy, limb girdle 2B	CS061275	DYSF	A	G	GCCACTC**A**CTCTGGC
Cockayne syndrome	CS099993	ERCC6	A	G	CTTTGCA**A**ACTCCTA
Multiple exostoses ?	CS068396	EXT1	C	A	CCCTCCC**C**ACTGCCT
Hypoprothrombinaemia	CS984089	F2	C	G	CCGTAGC**C**TCACTCC
Haemophilia A	CS076620	F8	A	G	CTGTCAG**A**CAACCAA
Haemophilia B	CS982186	F9	A	G/T	ACCGTTA**A**TTTGTCT
Haemophilia B ?	CS045815	F9	C	G	GCTGTTA**C**TGTCTAT
Fanconi anaemia	CS032696	FANCA	A	G	TGTTCTC**A**TTCTGTG
Contractural arachnodactyly	CS072199	FBN2	A	C	CATACTA**A**GATATTG
Contractural arachnodactyly	CS971736	FBN2	T	G	CACATAC**T**AAGATAT
Protoporphyria, erythropoietic	CS920753	FECH	C	T	TTTCATG**C**GAGCACT
Glycogen storage disease 2	CS971738	GAA	T	G	TCCCTCA**T**GAAGTCG
Thalassaemia beta	CS810003	HBB	G	A	GCCTATT**G**GTCTATT
Thalassaemia beta	CS001426	HBB	T	C	CTGCCTA**T**TGGTCTA
Sandhoff disease	CS890126	HEXB	G	A	TGCTTGC**G**GGGGGAT
Diabetes, MODY	CS083240	HNF4A	A	G	CCATCCA**A**CCATCCA
Glanzmann thrombasthenia	CS061294	ITGA2B	A	C	CCCTCTC**A**CCCTCAG
Long QT syndrome	CS094892	KCNH2	A	G	GGGGCTG**A**GCTCCCT
Fish eye disease	CS961608	LCAT	T	C	GCTGCCC**T**GACCCCT
Hypercholesterolaemia	CS961611	LDLR	C	T	CTCCTGG**C**GCTGATG

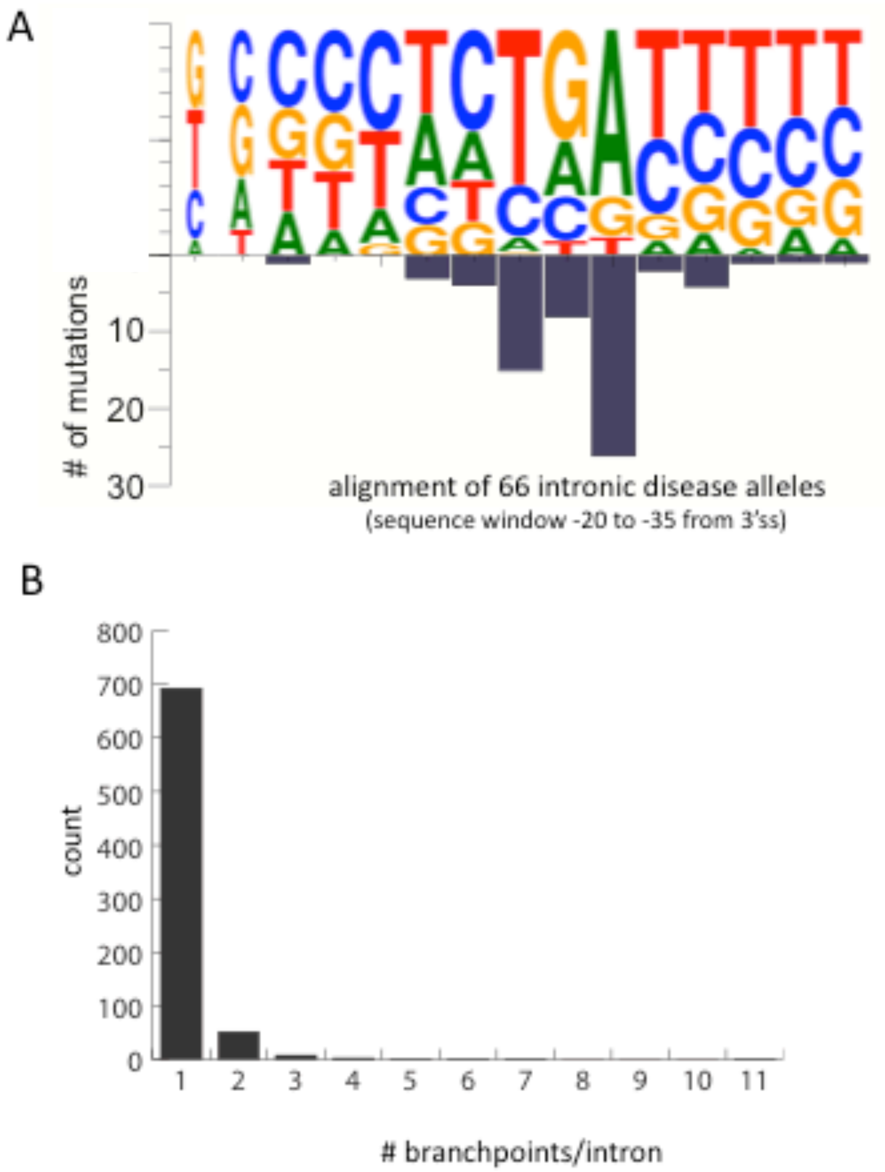
Mediterranean fever, familial	CS055595	MEFV	A	G	AAATTCA**A**GCTTTTC
Multiple endocrine neoplasia 1	CS067834	MEN1	C	A	GACCCTC**C**CTCCCCC
Cardiomyopathy, hypertrophic ?	CS041890	MYBPC3	C	A	AGCCTCA**C**TGGGGGT
Usher syndrome 1b	CS991465	MYO7A	G	A	GGCCTCT**G**ACATGCG
Neurofibromatosis 2 ?	CS942129	NF2	T	A	ACTTAGC**T**CCAATGA
Niemann-Pick disease C	CS043367	NPC1	A	G	TCCACTA**A**TGCTATT
Congenital disorder of glycosylation 1a	CS061318	PMM2	A	G	CATTCTA**A**GTGTTTT
Congenital disorder of glycosylation 1a	CS061319	PMM2	A	G	AGCCTTC**A**TCTGTAC
Protein C deficiency	CS952206	PROC	T	G	TGGCCGC**T**GACCCCC
Pancreatitis, chronic ?	CS066647	PRSS1	C	T	CTCCATA**C**AACTTGT
Retinoblastoma ?	CS063381	RB1	A	G	ATCCTCG**A**CATTGAT
Retinoblastoma	CS083264	RB1	A	C/G/T	ATTACTA**A**TTGGTAT
Brugada syndrome	CS994154	SCN5A	C	T	ACAAGGG**C**CTAATGC
Paraganglioma	CS013318	SDHD	T	C	GGTTTTT**T**ATTGATG
Cystinuria	CS050111	SLC3A1	C	G	AGGGTAA**C**CATGTCG
Pancreatitis, chronic ?	CS032084	SPINK1	A	T	GGAAATG**A**TTCTGTT
Extrapyramidal movement disorder	CS003079	TH	T	A	TCTGGGC**T**GATGCTG
Tuberous sclerosis	CS992717	TSC1	T	C	GTTGGTG**T**TCCTCAA
Porphyria, erythropoietic	CS100777	UROS	T	G	GGTGTGC**T**GAAGCCC
X-linked myopathy with excessive autophagy	CS092160	VMA21	A	C/T	GGTTCTG**A**TTTTCTC
Von Willebrand disease 1	CS070412	VWF	A	T	GCAAGTG**A**CCTCCTT
Thrombocytopaenia ?	CS102394	WAS	A	C	AGGATTC**A**CTGGAGT
Xeroderma pigmentosum (C)	CS040564	XPC	A	G	AGTGGAG**A**TAGAGAT

Supplementary Table 1 Human disease alleles that disrupt branchpoints.
We selected the 66 HGMD disease-causing splicing mutations that fell 20 to 35 bases upstream of a 3’ss for examination (62 unique mutation positions). The table lists the disease, HGMD ID, gene alias, wildtype allele, mutant allele and 15nt sequence window around the mutation. These sequences were aligned to create the branchpoint motif in Supplementary Figure 1.

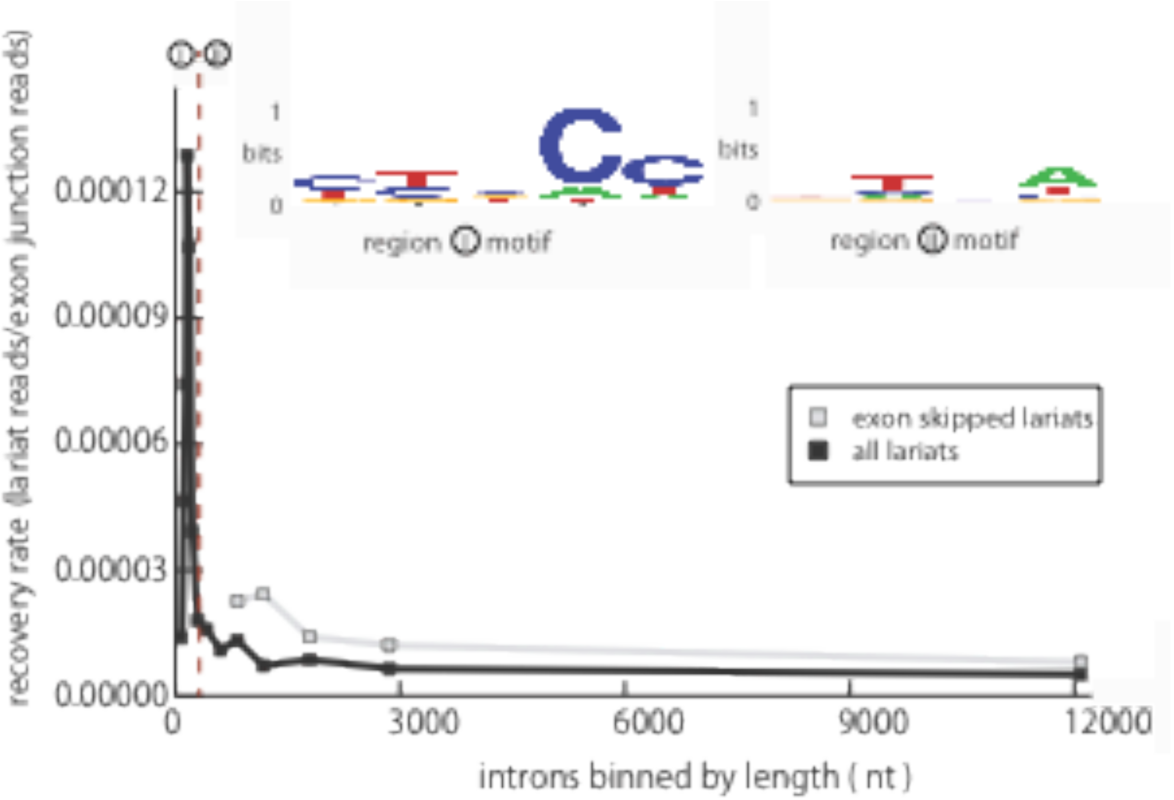
Supplementary Table 2

	BPS (Position)		
Sequence From	RT-PCR	Illumina Data	Gao et al., 2008
ACTB	CCAGTG (-29)	NA	CCAGTG (-29)
	TGTGAC (-25)		
	TGACAT (-23)		
	ACATGG (-21)		
	CAGATC (0)		
	CGT CTG (-57)		
CAD	TGCCAC (-29)	TGCCAC (-29)	NA
	GCCACC (-28)	GCCACC (-28)	
	CACTTG (-23)	TGCTGC (-32)	
EEF1A1	TGTGTT (-37)	TAACCA (-27)	AACGAC (-18)
	GGAGTT (-43)		
EIF4G3	TAATTT (-41)	GTAATT (-42)	NA
	GTAATT(-42)		
	TAATAG (-47)		
	TTTTAT (-36)		
EP300	CAG GTT(0)	CAGGTT (0)	NA
	CTTCCA (-3)		
FTCD	ACCCCT (-22)	TACCAG (-28)	NA
FUNDC2	TCATTT (-31)	CTCATT (-32)	NA
	AATAAC (-48)		
HSPG2	CTCCAC (-38)	CTCCAC (-38)	NA
KIAA0441	TGACAC (-24)	AAGGCC (0)	NA
	ACT GAC (-26)		
KIAA0913	CCCACC (-19)	GCCCAC (-20)	NA
MCM7		TGCTCT (-35)	NA
		TCCTGT (-41)	
		ATCCTG (-42)	
	CCATCC (-44)	CCATCC (-44)	
		CCCATC (-45)	
		TTCCCC (-48)	
	CTCCTT (-53)	CTCCTT (-53)	
	CTCTCC (-55)	CTCTCC (-55)	
	GCTCTC (-56)	GCTCTC (-56)	
		TGCTCT (-57)	
		CTCTGT (-62)	
PSMC3	TCCTTT (-27)	CTCCTT (-28)	NA
	CTCCTT (-28)		
	GCTCCT (-29)		
	GTCTGC (-52)		
RAB9A	GGTATA(-35)	GTA ATG (-22)	NA
	TATAAA (-33)		
	TGTGTA (-41)		
SEZ6L2	CAGGAT (0)	CAGGAT (0)	NA
VEZF1	AAGTTT (0)	AAGTTT (0)	NA
WDR74	AGCCCT (-28)	AGCCCT(-28)	NA
	GCCCTG (-27)		

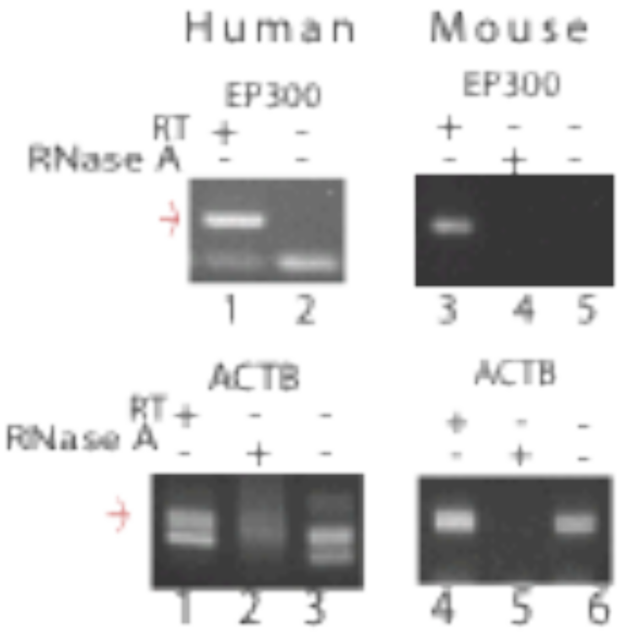
Supplementary Table 2. RT-PCR Validation of Illumina predicted branchpoints
Branchpoint predictions inferred from deep sequencing reads were compared to RT-PCR predictions and to a previous study of lariats in 20 human housekeeping genes[1]



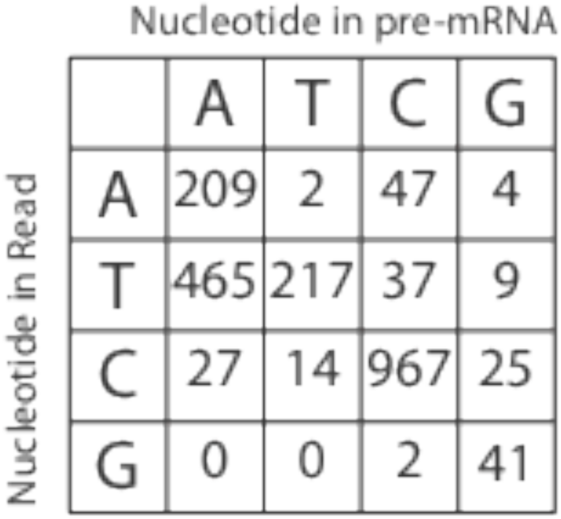
Supplementary Figure 1 Evidence for redundant and for non-redundant branchpoints in human pre-mRNA. A) Evidence for non-redundant branchpoints. The Human Gene Mutation Database was used to isolate splicing mutants that fall 20 - 35 nucleotides from the 3’ss, representing genetically defined non-redundant branchpoints. Alignment of sequences reveals branchpoint TRAY motif. The numbers of mutations at each position in the aligned motif are represented below. B) Evidence for redundant branchpoints: the distribution of branchpoints multiplicity per intron is represented by histogram covering 862 branchpoints in 760 introns. 9% of all introns sampled map to multiple lariat isoforms that utilize alternate branchpoints.



Supplementary Figure 2. Detection rate of lariats by intron length Only a fraction of the lariats created by splicing events were recovered (i.e. sampled by the Illumina deep sequencing data). Introns are binned by size and splicing events counted in the lariat and exon junction data. Exon junction data was used to estimate the expected lariat counts. The ratio demonstrates small introns are recovered as much as 10 times higher frequency than large introns. Branchpoint motifs from introns < 250 nucleotides contain a C rather than A. Intron turnover is initiated by debranching followed by rapid exonuclease digestion. As the rate limiting step in lariat turnover is debranching [2] and lariats that branch at C are debranched less efficiently [3], this result suggests that the initiations of turnover of small introns is more dependent on DBR1 than large introns.



Supplementary Figure 3. Conservation of intron circularization in mouse. Orthologous intron in mouse were assayed by lariat RT-PCR. . Red arrows represent lariat product sizes predicted from deep sequencing or human ortholog. PCR products were verified by sequencing. The RNA dependence of the amplification was verified by the omission of RT or pretreatment with RNAse A.



Nucleotide in pre-mRNA

Nucleotide in Read

	A	T	C	G
A	209	2	47	4
T	465	217	37	9
C	27	14	967	25
G	0	0	2	41

Supplementary Figure 4. Mutational profile of reverse transcriptase at branchpoints. Mutational matrix calculated from 2066 lariat/pre-mRNA alignments. A contingency table describes the correspondence between the sequence in the read and the sequence in the pre-mRNA as inferred from the reference genome. As branchpoint A is particularly mutable and often creates a transversion, distinct from the background mutational bias towards transitions, this mismatch represents a strong signal that can be used in future identification of lariat reads.

3) Supplementary Methods:

A) Computational Methods

I. Lariat Discovery

Overview Lariats were identified by inverted gapped alignments that mapped at the 5’ss and within 500 nt of 3’ss that had been annotated or observed in EST or deep sequencing data. Lariats outside of annotated sites were determined by scoring the quality of match to the 5’ss at the region in the read immediately downstream of the point of inversion (Figure 1A). Scoring *bona fide* lariats for matches to the 5’ss motif we find that 83% score greater than 6.0. Of inverted gapped intronic alignments that lack transcript support, 0.5% score greater than 6.0 - a two fold excess relative to non-inverted gapped intronic alignments (0.22%) or random intronic windows sampled (0.29%). Reverse transcriptase mutation rate is forty fold higher reading through a 2-5’ (vs 3-5’) phosphodiester bond (Supplementary Figure 2). Demanding this mutation at potential lariats increases this two fold excess to thirteen fold. From this high confidence set the proportion of each class of event is extrapolated to the whole data set.

Illumina Dataset.

To identify lariat branchpoints, we analyzed the Illumina Human Body Map 2.0 total RNA deep sequencing library. Reads consist of RNA samples derived from 16 human tissues. Most reads are 100 bp in length and all linker sequences were removed prior to analysis.

Hg19 Annotated.

We discovered branch points by searching for reads with non-canonical arrangements of intronic sequences. Reads that aligned to the hg19 genome with 3 or less mismatches were eliminated. Each remaining read was split into all possible head and tail segments, in which all heads and tails were at least 15 nt long. We mapped all head and tail segments to the hg19 genome using bowtie [4]. Head and tail segments were required to map to only one place in the genome and without any errors. Reads with segments that mapped in the expected order or did not map intronically were filtered from the dataset. The remaining inverted reads were mapped to splice sites. In cases where the read tail begins at the first nucleotide of the intron and the read head mapped within 500 nt upstream of a 3’ss, we determined that the read spans the lariat 2’-5’ linkage. In cases where there is alignment ambiguity, we allow up to two mutations and assume the alignment in which the tail maps to the first nucleotide of the intron. The last nucleotide of the read head was determined to be the branchpoint. 2066 lariat reads were discovered through this screen (lariat_0 – lariat_2065 in BED track). Reads that suggested branchpoints supported by spliced EST evidence were also reported. Nine lariat reads were discovered through this screen (lariat_2110 – lariat_2118 in BED track).

Unannotated, but with Illumina transcript support.

To discover lariats forming in transcripts that are unannotated in the hg19 assembly, we built a library of potential spliced products inferred from the inverted reads. For each inverted read that did not map to an annotated 5’ss, we constructed a potential upstream exon by taking an 85 nucleotide window immediately upstream of the read tail. We created an array of 200 potential downstream exons by taking 85 nucleotide windows at a distance of 1 to 200 nucleotides from the end of the read head. These windows were artificially spliced together to create a set of potential spliced products. We aligned the Illumina reads against these spliced products using bowtie, requiring that the read contained at least 15 nucleotides on either side of the splice junction and did not have any mismatches. In cases where a splice product was found and the implied intron contained a 5' GT and 3' AG sequence, the inverted read was determined to be a true lariat forming in an unannotated transcript. Forty-four additional lariat reads were discovered through this screen (lariat_2066 – lariat_2109 in BED track).

Lariats forming deep within introns.
The remaining out-of-order reads with intronic heads and tails, but without annotated or Illumina transcript support, were studied. From these reads, we filtered a high confidence set of lariats by requiring that both the heads and tails were never annotated as exons (alternative events), the beginning of the tail had a patser score of at least 6.0 against a 5'ss position specific weight matrix, and the read had a mutation at the branchpoint[5]. The 5'ss position specific weight matrix was created by inputting all hg19 annotated 5'ss sequences into the patser program. From within this high confidence set of internal lariats, the mutational profile was similar to the mutational profile of the bona-fide lariats (mostly A-> T mutations). We counted the number of splicing events that used an annotated 5'ss without a 3'ss, an annotated 3'ss without a 5'ss, or an event deep within an intron (using no annotated splice sites). We also counted the number of bona-fide lariats that passed the patser score, mutational, and intronic filters, and used that fraction to extrapolate how many true lariats without transcript support we expect exist in our data.

As a control, in-order reads that were most likely caused from template switching were passed through these same filters. There is a 3.3 fold increase of out-of-order reads that pass these filters compared to in-order reads, and the mutational profile of the in-order reads was different than bona-fide lariats, suggesting that these internal lariats are truly forming.

Identifying hereditary disease mutations in branchpoint motifs. We selected the 66 HGMD disease-causing splicing mutations that fell 20 to 35 bases upstream of a 3'ss for examination (62 unique mutation positions). These mutations plus seven nucleotides of flanking sequence on either side were used as input for a ClustalW multiple sequence alignment [6]. We used a maximal gap open penalty in order to align the sequences without gaps. The ClustalW output was used to create a sequence logo with the application WebLogo 3 [7]. We counted the number of times a mutation occurred at each position within the sequence logo to create a histogram showing how often a particular position within the sequence was affected by a splicing mutation.

II. Analysis

Branch Point Characterization.

Conservation. Introns with a minimum of 5 reads were separated into single branchpoint and multi-branchpoint categories. 51 introns were single branchpoint, 19 introns were multiclass containing a total of 71 branchpoints. Conservation of branchpoint motifs were estimated from mammalian phastCon score averaged over a 7 nt window centered on branchpoint. Single branchpoint introns had a higher level of conservation 0.229 versus 0.013. T-test was used to estimate significance: P value = 10^{-22}.

Distance. The branch point distance was measured as the distance between the last nucleotide of the read head to the first downstream annotated 3' splice site.

Mutational Profile: The mutational profile of reads that suggested a branch point at the last nucleotide of the intron (implying a circular intron) were compared to the mutational profile of all other reads with 'G' nucleotide branchpoint more distal from the 3'ss. A chi-squared test was used to show that these mutational profiles were significantly different.

Comparison of BP Distance in Alternative/ Constitutive Events: mRNA exon junction data was studied using tophat[8]. We created a junction file consisting of all possible constitutive and exon skipping events within each annotated transcript. We aligned the Illumina reads using the hg19 genome and this junction file to determine how many reads span each exon/exon junction. We calculated overall rates of alternative splicing and intersected this data with our lariat branchpoints that mapped to transcripts with alternative 3'ss or exon skipping alternative events. In cases where the lariat formed over a skipped exon or immediately upstream of a skipped exon, the branchpoint distance was measured to the first downstream exon. In cases where the lariat formed near an alternative 3'ss, the branchpoint distance was measured to the most upstream 3'ss. For all three classes of alternative events, p values were determined by randomly sampling the entire dataset 1000 times and counting how many times the average branchpoint distance was at least as extreme as the branchpoint distance in the alternative event.

Building Decision Tree to model 3'ss selection.
Overview:
The number of 'AG' dinucleotides between the branchpoint and the 3'ss 'AG' were counted for the 2066 hg19 annotated lariats. 2066 introns were selected at random, and simulated branchpoints were selected by randomly distributing the branchpoint distance distribution that was observed in real lariats. The number of 'AG' dinucleotides between these simulated branchpoints and the 3'ss 'AG' were counted. This process was repeated 1000 times to generate a p-value.

AG selection analysis was determined using the C5.0 software tool [9]. Lariat introns with exactly one branchpoint and one used 3'ss were considered in this analysis. The used AG, immediately upstream AG (if extant), and downstream AG were included in the dataset. First, the data was organized into a decision tree using classifiers from the literature, including distance to branchpoint, distance to upstream and downstream AGs, the nucleotide upstream of the AG, and presence of secondary structure (Gibbs free energy determined using RNAfold) [10]. In this initial run, the only informative classifiers were the presence or lack of an upstream AG, the distance between the AG and the branchpoint, and the distance to upstream and downstream AGs. Next, a range of constraints for each of the distance classifiers were applied to the dataset, and C5.0 was run on each of combination of classifier constraints. The classifier sets with error rates lower than the literature classifier sets were run again on the dataset, this time using half of the dataset as training data, and the other half as testing. The highest predictive scoring classifier sets were subjected to 10-fold cross validation trials. These cross-validation trials were completed 1000 times. The classifier set with the highest average predictive accuracy was used to create the decision tree.

RNA-protein Interactions.

Published CLIP data for FOX2 [11], PTB [12], hnRNP C [13] and a panel of other hnRNP proteins[14] were mapped around our branchpoint coordinates. The FOX2,PTB, hnRNP A1, hnRNP A2B1, hnRNP F, hnRNP M, and hnRNP U datasets were smoothed by using the center CLIP coordinate and adding 15 nucleotides to either side. The raw hnRNP C CLIP reads were aligned using bowtie and the last nucleotide was used as the binding point (as described in study). We smoothed the hnrRNP C data by adding 15 nucleotides to either side of the binding point. Of this set of proteins, we included CLIP tag density plots for the proteins with at least 1000 CLIP tag/lariat intron overlaps.

Lariat Recovery.

Lariats are sampled from a steady state population that is determined by their rate of splicing and their rate of degradation. Influences that alter the stability and sequence of the observed lariats are described in Supplementary Figure 1.

The number of reads spanning each annotated exon/exon junction was determined using tophat. Lariats and exon/exon junctions were both binned by intron size. The recovery rate of a lariat read was calculated by counting the number of detected lariat reads and dividing it by the number of detected exon/exon junction reads within each intron bin. The error bars were calculated by resampling the lariat read data 1000 times and using the 95% confidence interval.

III Statistical tests

Circular Introns have a significantly different mutational profile than conventionally located branchpoints in lariat introns - the $\chi 2$ test was used to compare the proportion of reads with an unambiguous base substitution at the transition site when this site occurred at position 0 versus other locations. In Figure 1B, the average branchpoint-3'ss distance was measured for 2066 reads. Resampling was used to compare the average distance of a series of 1000 samples from different subset of alternative exons from the listed categories to the measured value. To determine the significance of 80% of the 3'ss being the first "AG" downstream of the splice site, permutation trials simulated branchpoints (maintaining average 3'ss/branchpoint distance) in introns to scan for first "AG".

B) Experimental Methods

Overview of Validation We performed nested RT-PCR to validate the branchpoint predictions of lariats in total RNA from HEK293 cells. All sequences (primer, PCR), experimental or computational protocols and statistical tests are available at http:/fairbrother.biomed.brown.edu/data/Lariat

Analysis of Branch Points by Nested RT-PCR

We performed RT-PCR analysis to generate an amplified product spanning the branchpoint of lariat splicing intermediates. HEK293 cells were grown to 80-90% confluence in

DMEM+10%FBS supplemented with 1000 U/ml of each penicillin and streptomycin (Gibco Cat#15140-122) at 37 ^{O}C in 5% CO_2. Total RNA was isolated using Trizol reagent (Invitrogen, Cat#15596-026) according to the manufacturer's instructions with minor modifications. Following the initial precipitation of RNA, samples were digested with Turbo DNase (Ambion, Cat#AM2239) for 10 min at 37^{O}C to remove any DNA contamination. Samples were then extracted twice with 1:1 phenol:choloroform, pH 4.5 and twice with chloroform. RNA was then washed according to the manufacturer's instructions and re-suspended in DEPC-treated H_2O. cDNA was synthesized using random 9-mer primers (Integrated DNA Technologies) and Superscript-III Reverse Transcriptase (Invitrogen, Cat# 18080-093) according to the manufacturer's instructions. PCR was then performed on 0.5µl of cDNA using the "outer" primer pair (sequences available at http://fairbrother.biomed.brown.edu/data/Lariat) with 0.5U Platinum Taq DNA Polymerase (Invitrogen, Cat#10966-018) in a total volume of 25µl according to the manufacturer's instructions. If necessary, 1µl 20mg/ml RNase A was added to the PCR reactions and samples were incubated at 37^{O}C for 30min before running the PCR. A second PCR using "nested" primers (sequences available at http://fairbrother.biomed.brown.edu/data/Lariat) was then performed with 0.5µl of the initial PCR product used as the template. It was necessary to optimize conditions separately for each reaction. Exact conditions are available upon request. The product of the second PCR was then separated on a 2% agarose gel and the appropriate bands were excised and purified using a Quiagen gel extraction kit (Quiagen, Cat#28704). PCR products were then cloned into pCR2.1 using a TOPO TA Cloning Kit (Invitrogen, Cat#45-0641) and transformed into TOP10 *E. coli* cells. Individual colonies were then grown in LB+ampicillin and plasmid DNA was isolated using a Quiagen Miniprep Kit (Quiagen, Cat#27106) and was subsequently sequenced.

1. Gao, K., Masuda, A., Matsuura, T. & Ohno, K. Human branch point consensus sequence is yUnAy. *Nucleic Acids Res* **36**, 2257-67 (2008).
2. Nam, K., Lee, G., Trambley, J., Devine, S.E. & Boeke, J.D. Severe growth defect in a Schizosaccharomyces pombe mutant defective in intron lariat degradation. *Mol Cell Biol* **17**, 809-18 (1997).
3. Nam, K. et al. Yeast lariat debranching enzyme. Substrate and sequence specificity. *J Biol Chem* **269**, 20613-21 (1994).
4. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S.L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* **10**, R25 (2009).
5. Hertz, G.Z. & Stormo, G.D. Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. *Bioinformatics* **15**, 563-577 (1999).
6. Chenna, R. et al. Multiple sequence alignment with the Clustal series of programs. *Nucl. Acids Res.* **31**, 3497-3500 (2003).
7. Crooks, G.E., Hon, G., Chandonia, J.M. & Brenner, S.E. WebLogo: a sequence logo generator. *Genome Res* **14**, 1188-90 (2004).
8. Trapnell, C., Pachter, L. & Salzberg, S.L. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* **25**, 1105-1111 (2009).
9. Quinlan, R.S. Induction of Decision Trees. *Mach. Learn.* **1**
, 81-106 (1986).

10. Ding, Y., Chan, C.Y. & Lawrence, C.E. Sfold web server for statistical folding and rational design of nucleic acids. *Nucl. Acids Res.* **32**, W135-141 (2004).
11. Yeo, G.W. et al. An RNA code for the FOX2 splicing regulator revealed by mapping RNA-protein interactions in stem cells. *Nat Struct Mol Biol* (2009).
12. Xue, Y. et al. Genome-wide analysis of PTB-RNA interactions reveals a strategy used by the general splicing repressor to modulate exon inclusion or skipping. *Mol Cell* **36**, 996-1006 (2009).
13. Konig, J. et al. iCLIP reveals the function of hnRNP particles in splicing at individual nucleotide resolution. *Nat Struct Mol Biol* **17**, 909-15.
14. Huelga, S.C. et al. Integrative Genome-wide Analysis Reveals Cooperative Regulation of Alternative Splicing by hnRNP Proteins. **1**, 167-178.