10564–10578 *Nucleic Acids Research, 2014, Vol. 42, No. 16* *Published online 14 August 2014*
doi: 10.1093/nar/gku744

A comprehensive survey of non-canonical splice sites in the human transcriptome

Guillermo E. Parada[†], Roberto Munita[*][†], Cledi A. Cerda and Katia Gysling[*]

Nucleus Millennium in Stress and Addiction, Department of Cellular and Molecular Biology, Faculty of Biological Sciences, Pontificia Universidad Católica de Chile, Alameda 340, Santiago, Chile

Received June 11, 2014; Revised July 31, 2014; Accepted August 1, 2014

ABSTRACT

We uncovered the diversity of non-canonical splice sites at the human transcriptome using deep transcriptome profiling. We mapped a total of 3.7 billion human RNA-seq reads and developed a set of stringent filters to avoid false non-canonical splice site detections. We identified 184 splice sites with non-canonical dinucleotides and U2/U12-like consensus sequences. We selected 10 of the herein identified U2/U12-like non-canonical splice site events and successfully validated 9 of them via reverse transcriptase-polymerase chain reaction and Sanger sequencing. Analyses of the 184 U2/U12-like non-canonical splice sites indicate that 51% of them are not annotated in GENCODE. In addition, 28% of them are conserved in mouse and 76% are involved in alternative splicing events, some of them with tissue-specific alternative splicing patterns. Interestingly, our analysis identified some U2/U12-like non-canonical splice sites that are converted into canonical splice sites by RNA A-to-I editing. Moreover, the U2/U12-like non-canonical splice sites have a differential distribution of splicing regulatory sequences, which may contribute to their recognition and regulation. Our analysis provides a high-confidence group of U2/U12-like non-canonical splice sites, which exhibit distinctive features among the total human splice sites.

INTRODUCTION

Most genes in higher eukaryotes are interrupted by non-coding sequences, called introns, which are precisely excised from pre-mRNAs during splicing. Nuclear pre-mRNA introns are processed by the spliceosome, a complex macromolecular machine composed of five small nuclear RNAs and numerous proteins (1,2).

Proper intron recognition and removal rely on consensus sequences located at the intron/exon boundaries. Dinucleotide sequences at these boundaries have been found to be strongly conserved and relevant for proper splicing (3–5). Nearly all introns belong to the so-called U2-type, which are spliced by the major spliceosome and are flanked by GT–AG splice site dinucleotides. The most frequent exception to this rule are the U2-type GC–AG splice sites, comprising ∼0.9% of human splice sites (6). On the other hand, about 0.4% of the human splice sites belong to the U12-type. These introns are processed by the minor spliceosome and even though they were first described to have AT–AC dinucleotides at the intron/exon boundaries, the vast majority of them contain GT–AG sites (7). Indeed, the AT–AC sites comprise only ∼0.09% of the splice sites (6).

Despite the disruptive splicing effects that have mutations of splice site dinucleotides (3–5), introns with non-canonical splice sites (that is, with sequences other than GT–AG, GC–AG or AT–AC at the intron/exon boundaries) have been reported to be efficiently removed (6,8–12). These reported non-canonical splice sites have U2/U12-like splice site consensus sequences (U2/U12-like non-canonical splice sites). For instance, evolutionary conserved U2-like introns with GA–AG splice sites have been identified in FGFR genes (8,9) and a functional GT–TG splice site has been found in the GNAS gene (10,11). Although the first global analysis of splice sites in the human transcriptome, conducted 14 years ago, did not find confident evidence for non-canonical splice sites (13), most recent analyses based on expressed sequence tag (EST) sequences have reported U12-like non-canonical splice sites (6) and more examples of U2-like GT–TG introns (12).

The advent of high-throughput sequencing technologies has provided an unprecedented opportunity to explore the complexity of mammalian transcriptomes (14). For instance, analyses of RNA-seq data have resulted in the discovery of thousands of new splice sites and alternative splicing events in the human transcriptome (15–17). However, the high resolution power of high-throughput sequencing

[*]To whom correspondence should be addressed. Tel: +562 23542654; Fax: +562 23542660; Email: kgysling@bio.puc.cl
Correspondence may also be addressed to Roberto Munita. Tel: +562 23542657; Fax: +562 23542660; Email: rnmunita@uc.cl
[†]The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

has not been used to generate a non-canonical splice site catalog on the human transcriptome.

To make a comprehensive analysis of non-canonical splice sites present in the human transcriptome, we have processed nearly 3.7 billion RNA-seq reads from 16 human tissues and a lymphoblastoid human cell line (GM12878). Our systematic analysis provides a list of high-confidence non-canonical splice sites and an insight into their characteristic features. Our comprehensive identification of non-canonical splice sites will improve the human transcriptome annotation. Further understanding of the mechanism underlying the recognition and processing of non-canonical splice sites could expand our knowledge of the splicing process. We provide the full annotation and quantification of the entire list of high-confidence canonical and non-canonical splice junctions for each analyzed human tissue (available as a UCSC Hub at http://54.214.245.35/Tracks/Splicing/hub.txt).

MATERIALS AND METHODS

Processing of RNA-seq data

We used the used RNA-seq data of GM12878 cell line provided by ENCODE project (18) and RNA-seq data of a mixture of 16 human tissues generated by Illumina Body Map 2.0 project (for additional information see Supplementary Data). The reads were processed in order to remove the adapters and low-quality sequences (PHRED score ≤ 10) with FASTX toolkit (http://hannonlab.cshl.edu/fastx_toolkit/index.html) and sickle (https://github.com/najoshi/sickle) for trimming of paired-end reads. After this, only the reads ≥50 nt long were kept. RNA-seq reads of GM12878 provided from the ENCODE project (18) (Supplementary Data) were aligned to the diploid genome of GM12878 (19) and the RNA-seq of the 16 human tissues mixture to the reference genome (hg19) using MapSplice (20). MapSplice was configured to detect canonical and non-canonical splice junctions with anchor length ≥8, a minimal intron size of 1 nt and allowing two mismatches 25 nt each.

Ab initio detection of splice junctions

Splice junctions were extracted from unique gapped alignments that have an anchor ≥8 nt. To avoid false non-canonical introns derived from alignment errors, the read alignments that have non-canonical splice junctions were corroborated by BLAT (21) and were discarded if they had repetitive sequences present at Repbase (22) or low complexity sequences detected by DUST (23). A splice junction was considered present in a dataset of RNA-seq data if at least three different sequences supported the splice junction. GM12878 RNA-seq data were aligned to their paternal and maternal genome (diploid genome). We only considered those splice junctions that were in at least three coincident alignments in maternal and paternal GM12878 genomes.

We extracted splice junctions from cDNA and EST alignments provided by UCSC Genome Browser database (24). We removed cDNA/EST sequences that were repetitive or had low complexity (using Repbase and DUST). Splice junctions were extracted from unique gapped alignments that had an anchor ≥15 without mismatches. The splice junctions that were supported by three or more cDNA/EST sequences were considered as present in cDNA/EST alignments.

As the alignments to the reference genome can span over indels and single nucleotide polymorphisms (SNPs), false non-canonical splice junctions could be detected. For this reason, we discarded the gapped alignments to the reference genome which had non-canonical splice junctions that spanned over indels or SNPs reported on SNPdb135.

Library-guided detection of splice junctions

All the splice junctions identified in RNA-seq and cDNA/EST alignments were used to create a library of splice junctions (Figure 1A). We mapped the previously used RNA-seq data, in addition to tissue-specific RNA-seq data from 16 human tissues, to the library of splice junctions using Bowtie (25), configured to align with a SOAP-like '-v 2' mode allowing two mismatches with a seed length of 50 nt (Figure 1B). We checked that the new splice junction mapping reads found did not map to other genomic location and did not have repetitive or low complexity sequences. Moreover, to ensure the reproducibility of the splice junction detection, we selected the splice junctions that were present in at least two different data sources (tissues extracted from the same individual were considered as one data source) (Figure 1C).

Non-canonical splice junction processing

Direct repeat sequences located at the splice junctions generate alignment ambiguity and multiple putative non-canonical splice sites. This phenomenon can lead to false non-canonical splice site annotation (13). For this reason, when a non-canonical splice junction was flanked by direct repeats, we checked that it did not have ambiguous canonical splice sites (13). Moreover, based on the position weight matrices (PWMs) reported for canonical human splice junctions (6), we selected the alignment that yields the non-canonical splice junction with the best fit into the canonical PWMs.

Based on the canonical splice junction PWMs, we classified the non-canonical introns as U2/U12-like or non-U2/U12. To do this, we assigned scores based on the PWMs to a population of splice junctions with scrambled sequences. We took the 95th percentile of the scrambled splice junction scores as the U2/U12 threshold score to classify non-canonical splice junction as U2/U12-like. If a non-canonical splice junction had a score higher than 70.00, it was classified directly as U2/U12-like. The non-canonical splice junctions that have scores between 70.00 and the threshold were classified as U2/U12-like only if they shared a splice site with a canonical splice junction. Finally, if the non-canonical splice junction had a score below the threshold score or its intron was smaller than 80 nt, it was classified as non-U2/U12.

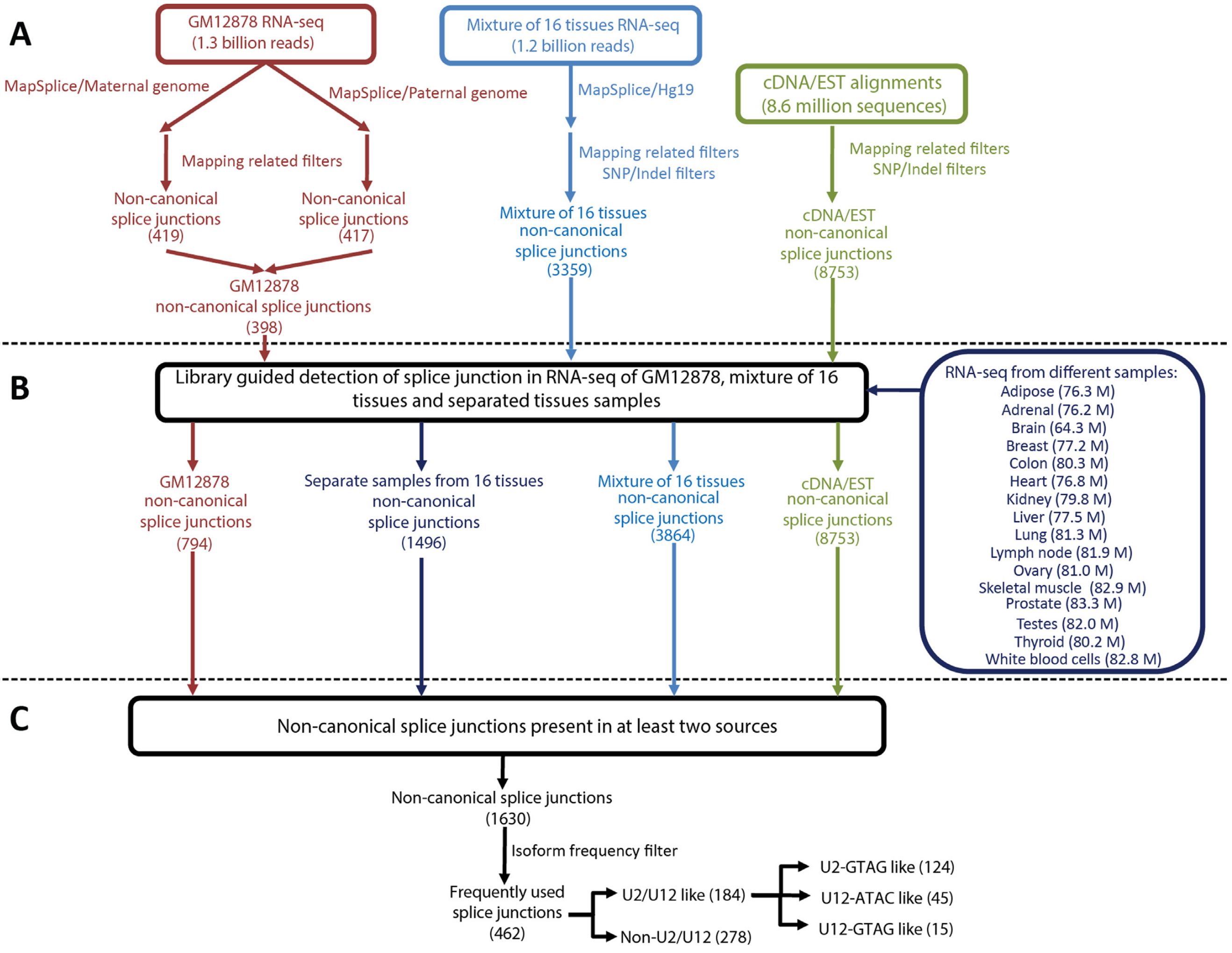

Figure 1. An overview of the workflow used for the search of non-canonical splice junctions. (**A**) *Ab initio* detection of non-canonical splice junctions. RNA-seq data of GM12878 were aligned to their diploid personalized genome and RNA-seq data from a mixture of 16 human tissues were aligned to the human reference genome (hg19) using MapSplice. All alignments of RNA-seq and cDNA/EST data were pre-processed in order to generate an initial library of splice junctions. Additional SNP/indel filters were applied to tissue RNA-seq and cDNA/EST alignment. (**B**) All the RNA-seq data were re-aligned to the library of splice junctions. Additional RNA-seq data from individual tissues were also directly aligned to the library. (**C**) A total of 1630 non-canonical introns were present in at least two sources of data. From these, 462 non-canonical splice junctions were detected in at least a coverage ratio of 1:20 compared with their most abundant splice variant.

Alternative splice site analysis

After computing all the 5′ and 3′ splice sites identified, we classified the alternative splice site events in exon skipping, exon inclusion, intron retention and 5′/3′ alternative splice sites. Based on the total RNA-seq coverage of each splice junction, we estimated non-canonical intron usage percentages (ψ). To select the non-canonical splice junctions that could have functional relevance, we filtered the alternative non-canonical splice junctions. We only continue working with the non-canonical splice junctions that were detected in at least a coverage ratio of 1:20 compared with the predominant splicing variant.

Splice junction annotation as a UCSC Genome Browser track

We annotated the splice junctions found as a UCSC Genome Browser track (available as a UCSC Hub at http://54.214.245.35/Tracks/Splicing/hub.txt). The splice junctions are represented as an intron flanked by two exons of 8 nt. The ID of each splice junction encodes its splice site dinucleotides and read coverage. U2/U12-like introns are in a green scale, where lighter green colors indicate higher similitude with U2/U12 splice junctions. Non-U2/U12 introns are in a red scale, where lighter colors indicate higher differences with U2/U12 splice junctions. Moreover, all detected non-canonical splice sites are reported with their 0-based genomic coordinates (Supplementary Data).

Conservational analysis of the splice junctions

We extracted the splice junctions from mouse cDNA/EST alignments provided by the UCSC database (24) using the same processing protocol previously described for human cDNA/EST alignments. To compare the mouse and human splice junctions, we use LiftOver utility (24) to translate mouse genome coordinates to human genome coordinates. The splice junction was considered conserved if it had the same coordinates and splice site dinucleotides.

Analysis of splicing regulatory sequence distribution

We map the hexamers associated with splicing regulatory elements (SREs) that were identified by fluorescence-activated screening (FAS) (26–29) to the human genome. The SRE hexamers were grouped by EIEs (ESE + ISS) and IIEs (ISE + ESS) (30). We took 5′/3′ exonic and intronic windows of 100 nt long for all the splice junctions identified. We compared the density of SRE hexamers (number of SRE standardized by length of the genome area analyzed) in the different windows between canonical and non-canonical splice junctions. The significance of the differences found was assessed by Pearson Chi-squared test.

Experimental validation of non-canonical splice junctions by RT-PCR/Sanger sequencing

We used reverse transcriptase-polymerase chain reaction (RT-PCR) to validate 10 non-canonical U2/12-like and 4 non-U2/U12 splice junctions. We used Human Universal RNA samples (SABiosciences) and different RTs: MMLV (RevertAidFirst Strand cDNA Synthesis), AMV (Promega) or Transcriptor (Roche) to do the RT-PCR. The PCR amplification was carried out using a touchdown thermocycling program (31) during 35 cycles and products were analyzed by gel electrophoresis. The presence of the different isoforms detected were corroborated using Sanger sequencing of the RT-PCR products subcloned using the pGEM-T Easy Vector System (Promega). The primers for the RT-PCR are listed in Supplementary Table S1.

Secondary RNA structure prediction

In silico prediction of secondary RNA structure associated with non-U2/U12 splice junction was performed using RNA fold web server (32). The visualization of the RNA structures was obtained using PseudoViewer3 (33) and edited with Adobe Illustrator.

RESULTS

Using public transcriptome data to find high-confidence splice junctions

We have done a comprehensive search of canonical and non-canonical splice sites at the human transcriptome using public RNA-seq data and cDNA/EST data. Mapping RNA-seq data coming from different genomic backgrounds to the human reference genome can lead to false non-canonical splice junction detection due to indels in the reference genome or SNPs at the splice sites; thus we used the transcriptomic and genomic data from the same cell line (GM12878). Both sources of sequences derive from the same genomic background and, therefore, with the same genomic variations. The GM12878 is a lymphoblastoid cell line that has been deeply sequenced with both parents (as a trio) and ensembled as diploid genome, with maternal and paternal allele specific genomic variations (19). Using MapSplice (20), we first mapped 1.3 billion of RNA-seq reads from GM12878 (provided by the ENCODE project (18)) to the diploid genome of GM12878 and obtained non-canonical splice junctions free of genome variation associated errors (Figure 1A). Then, to get a comprehensive landscape of non-canonical splice sites at the different human tissues, we also analyzed the alignments of 1.2 billion RNA-seq reads from a mixture of 16 human tissues (Illumina Body Map 2.0) to the human reference genome and cDNA/EST alignments, but we filter non-canonical splice sites that have SNPs or indels reported in SNPdb135 (34) (Figure 1A).

Although MapSplice has shown high performance to discover novel splice junctions, *ab initio* detection of splice junctions is a computationally challenging task prone to sensitivity problems (35–37). In order to increase the sensitivity to identify RNA-seq reads coming from splice junctions, we directly mapped a total of 3.7 billion human RNA-seq reads from GM12878 (1.3 billion), a mixture of 16 human tissues (1.2 billion) and RNA-seq from individual tissue samples (1.2 billion) to the splice junction library of all introns that we initially found (Figure 1B). By doing this, we increased the number of reads that align to non-canonical splice junctions. We found 1630 non-canonical splice sites present in at least two different sources of data (Supplementary Data). For the following analysis, we only considered the 462 non-canonical sites that were found to have at least a coverage ratio of 1:20 compared with their most abundant splice variant (Figure 1C). We identified 278 non-U2/U12 non-canonical splice sites (Figure 1C). Non-U2/U12 non-canonical splice sites were analyzed separately because they could be attributed to sequencing errors and reverse transcriptase template switching events.

Diversity of U2/U12-like non-canonical splice sites

We found 184 high-confidence non-canonical introns that have U2/U12-like splicing sequences (Figure 1C, Table 1 and Supplementary Data), of which 51% are not annotated in GENCODE v17. Thus, this represents a group of introns rich in novel and unexplored splice junctions present in the human transcriptome.

We validated 9 of 10 selected U2/U12-like non-canonical introns by RT-PCR using a human RNA sample representative of many major human tissues (Supplementary Figures S1–S9). The sequences obtained by Sanger method were aligned to the human reference genome showing the same non-canonical introns identified herein.

Evolutionary conservation of U2/U12-like non-canonical splice sites

To assess the evolutionary conservation of U2/U12-like non-canonical splice junctions, we verified their presence in mouse cDNA/EST alignments. We found that at least 28% of identified human non-canonical splice sites are conserved in mouse. Moreover, there are non-canonical splice sites conserved across large evolutionary distances. For example, the GT–GG splice sites of BANP gene show a transversal conservation among almost all vertebrates, but in fishes it is processed as GT–AG canonical splice sites, because the 3′ GG splice site is a 3′ AG in the fish genome (Figure 2A).

Table 1. Diversity of human splice sites

Total splice sites			U2-like splice sites			U12-like splice sites		
Dinucleotides	**N°**	**%**	**Dinucleotides**	**N°**	**%**	**Dinucleotides**	**N°**	**%**
GTAG	219888	98.9297	**GTAG**	218584	99.0583	**GTAG**	1266	82.6911
GCAG	1976	0.8890	**GCAG**	1954	0.8855	**ATAC**	201	13.1287
ATAC	219	0.0985	**GTTG**	36	0.0163	**ATAG**	20	1.3063
GTTG	40	0.0180	**GGAG**	32	0.0145	**ATAA**	17	1.1104
GGAG	32	0.0144	**GAAG**	14	0.0063	**ATAT**	10	0.6532
ATAG	26	0.0117	**GTGG**	12	0.0054	**GTAT**	5	0.3266
ATAA	17	0.0076	**TTAG**	10	0.0045	**GCAG**	4	0.2613
GTGG	15	0.0067	**GTAA**	9	0.0041	**GTTG**	4	0.2613
GAAG	14	0.0063	**ATAG**	6	0.0027	**GTGG**	3	0.1960
ATAT	10	0.0045	**GTAT**	2	0.0009	**GTCG**	1	0.0653
TTAG	10	0.0045	**GTGT**	1	0.0005			
GTAA	9	0.0040	**GTGA**	1	0.0005			
GTAT	7	0.0031	**GTCC**	1	0.0005			
GTGA	1	0.0004						
GTCG	1	0.0004						
GTCC	1	0.0004						
GTGT	1	0.0004						

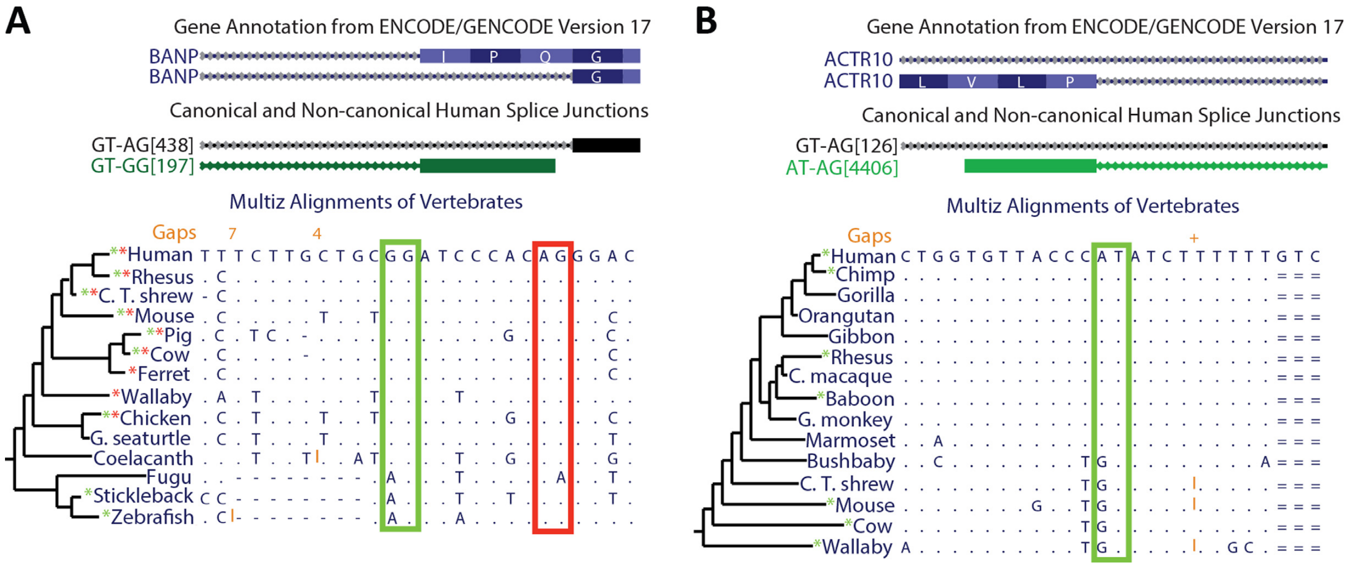
A
Gene Annotation from ENCODE/GENCODE Version 17
BANP
BANP
I P Q G
G
Canonical and Non-canonical Human Splice Junctions
GT-AG[438]
GT-GG[197]
Multiz Alignments of Vertebrates
Gaps 7 4
Human T T T C T T G C T G C G G A T C C C A C A G G G A C
Rhesus
C. T. shrew
Mouse
Pig
Cow
Ferret
Wallaby
Chicken
G. seaturtle
Coelacanth
Fugu
Stickleback
Zebrafish
B
Gene Annotation from ENCODE/GENCODE Version 17
ACTR10
ACTR10
L V L P
Canonical and Non-canonical Human Splice Junctions
GT-AG[126]
AT-AG[4406]
Multiz Alignments of Vertebrates
Gaps +
Human C T G G T G T T A C C C A T A T C T T T T T T G T C
Chimp
Gorilla
Orangutan
Gibbon
Rhesus
C. macaque
Baboon
G. monkey
Marmoset
Bushbaby
C. T. shrew
Mouse
Cow
Wallaby

Figure 2. Non-canonical splice site conservation across different vertebrates. The UCSC Genome Browser images show two examples of non-canonical splice sites that are conserved in several vertebrates. Based on our results, we made an annotation of human splice junctions (middle track). The splice junctions' name (ID) indicates their dinucleotides and read coverage (for details see the Materials and Methods section). Genome alignments from different vertebrates are shown. Dots indicate conservation of the human nucleotides; red and green rectangles indicate the conservation of canonical and non-canonical splice sites respectively; and * indicates that the splice site is supported by cDNA/EST alignments of each vertebrate. (**A**) The BANP gene has an alternative GG-3′ non-canonical splice site that is conserved among most of the vertebrates and it was derived from an ancient canonical splice site. (**B**) The AT–AG non-canonical splice site of ACTR10 has been recently derived in evolution from a canonical splice site. This non-canonical splice site is exclusive of some primates.

In contrast, the AT–AG non-canonical splice sites present in ACTR10 gene are only conserved in some primates because the 5′ AT splice site is a 5′ GT in other mammalian genomes (Figure 2B).

Most of U2/U12-like non-canonical splices are involved in alternative splicing

Our analysis shows that 76% of U2/U12-like non-canonical splice sites are involved in alternative splicing events, while 46% of canonical splice sites are involved in alternative splicing events. Moreover, we found that alternative 5′/3′ splice sites are the most abundant kind of alternative splicing events associated with U2/U12-like non-canonical splice sites (Figure 3A), consistent with previous reports (8,12).

U2/U12-like non-canonical splice site involvement in tissue-specific alternative splicing

Some introns that have U2/U12-like non-canonical splice sites are alternatively spliced with tissue-specific patterns. One example is a GT–TG splice site located at the TSPYL2 gene. The TSYPL2 gene encodes a protein that is part of the CASK/TRB1/TSPYL2 transcriptional complex which modulates gene expression in response to neuronal synaptic activity, probably by using its nucleosome assembly protein-like (NAP-like) domain to modulate nucleosome assembly (38). The NAP-like domain is interrupted by a premature stop codon when an alternative TG 3′ splice site is used. In tissues like adrenal and prostate, the TG 3′ splice site is used in more than 40% of the transcripts, but, in contrast, in brain and skeletal muscle, the TG 3′ splice site is used in less than 10% of the transcripts (Figure 3B).

We found non-canonical splice sites that are highly tissue specific. For example, the CPSF3 gene, which encodes a key component of the cleavage and polyadenylation factor complex, has a GG–AG non-canonical splice site that is only used on testis transcripts (Figure 3C). This non-canonical splice site is currently not annotated in GENCODE v17, probably due to its high tissue-specific pattern.

U2/U12-like non-canonical splice sites show differences in the abundance of splicing regulatory sequences

Cis-acting SREs are critical for splicing modulation (39). Several studies have identified exonic splicing enhancers (ESEs) (26), exonic splicing silencers (ESSs) (27), intronic splicing enhancers (ISEs) (28) and intronic splicing silencers (ISSs) (29) by FAS (40). To determine if there are different SRE patterns between canonical and non-canonical splice sites, we mapped hexamers enriched in the SRE sequences that were identified by FAS to the human reference genome and counted their occurrences around all identified splice sites.

In order to analyze how these different SRE patterns influence exon/intron recognition, we classified the ESE/ISS sequences as exon-identity elements (EIEs) and ISE/ESS sequences as intron-identity elements (IIEs) (30), and calculated their positional density within 5′/3′ exonic and intronic windows 100 nt long (Figure 4). As all of U2/U12-like non-canonical splice sites have only one non-canonical dinucleotide, we classified them into 5′ and 3′ non-canonical splice sites. Comparing EIE/IIE densities in exon–intron boundaries of canonical and non-canonical splice sites, we found that all significant differences indicate an overall enrichment of SREs along exon–intron boundaries that have non-canonical splice sites (Figure 4 and Supplementary Table S2). The same significant differences were observed when EIE/IIE densities in exon–intron boundaries of non-canonical splice sites were compared with EIE/IIE densities from alternative canonical splice sites (Supplementary Table S3).

Table 2. Edited U2/U12-like non-canonical splice sites

Gene	Intron type	Edited dinucleotides	Reference
NRK	U2-like	GT–**AA** > GT–**AG**	
ADARB1	U2-like	GT–**AA** > GT–**AG**	(41)
HOOK3	U2-like	GT–**AA** > GT–**AG**	
COG8	U2-like	GT–**AA** > GT–**AG**	
C14orf37	U2-like	**AT**–AG > **G**T–AG	
NUP210	U12-like	**AT**–AG > **G**T–AG	
CRYZL1	U12-like	**AT**–AG > **G**T–AG	

Some non-canonical splice sites are modified by RNA editing

Alternative splicing could be influenced by RNA editing. In particular, A-to-I editing was found to directly modify three reported cases of non-canonical splice sites (41–43). Because inosine is recognized as guanine by the spliceosome, alternative 5′-AT or AA-3′ can be turned into canonical splice sites through A-to-I editing. Since inosine is base-paired with cytosine during reverse transcription, sequencing machines also identify inosine as guanosine. We found three AT–AG and four GT–AA non-canonical splice sites that have consistent A>G mismatches in poly(A)-minus RNA-seq alignments from ENCODE cell lines (Table 2). These splice sites are involved in alternative splicing. For example, when the GT–AA non-canonical splice site of NRK gene is edited to GT–AI, a 42 nucleotide exon is included, making this an A-to-I editing dependent exon inclusion event (Figure 5).

Non-canonical introns as a source of transcript annotation errors

The existence of non-canonical introns is underestimated by many widely used RNA-seq mappers (44–49). For this reason, non-canonical introns are a frequent source of transcriptome misannotations. Some misannotations are due to the presence of putative GT–AG canonical splice sites near the real non-canonical splice sites. For example, the reads of a GA–AG non-canonical splice junction in the ITPR1 gene can be misaligned (with three mismatches) to a false canonical GT–AG (Figure 6A). This artifact is produced because

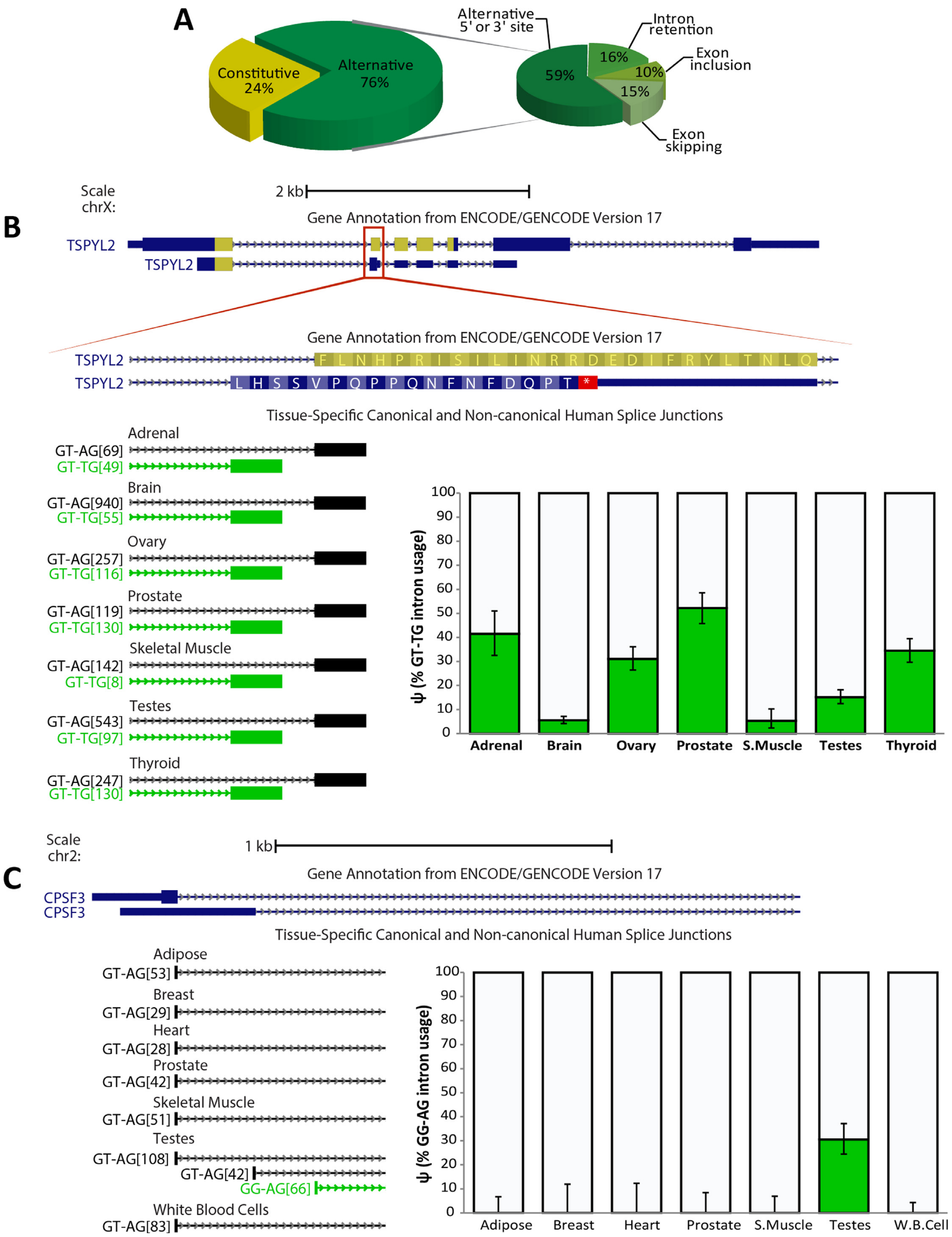

Figure 3. Non-canonical splice sites are highly involved in alternative splicing and some show tissue-specific alternative splicing patterns. (**A**) Participation of non-canonical splice sites in alternative splicing. (**B**) The use of a non-canonical splice site generates a frame shift and a premature termination codon that disrupts a NAP-like domain (highlighted in yellow) of the TSPYL2 protein. A tissue-specific pattern of 3′ alternative splice site selection is shown across seven human tissues. Coverage quantification is plotted, where error bars indicate the 95% binomial confidence interval. (**C**) CPSF3 gene has an alternative 5′-GG non-canonical splice site. Coverage quantification shows a testis-specific selection of the 5′-GG non-canonical splice site.

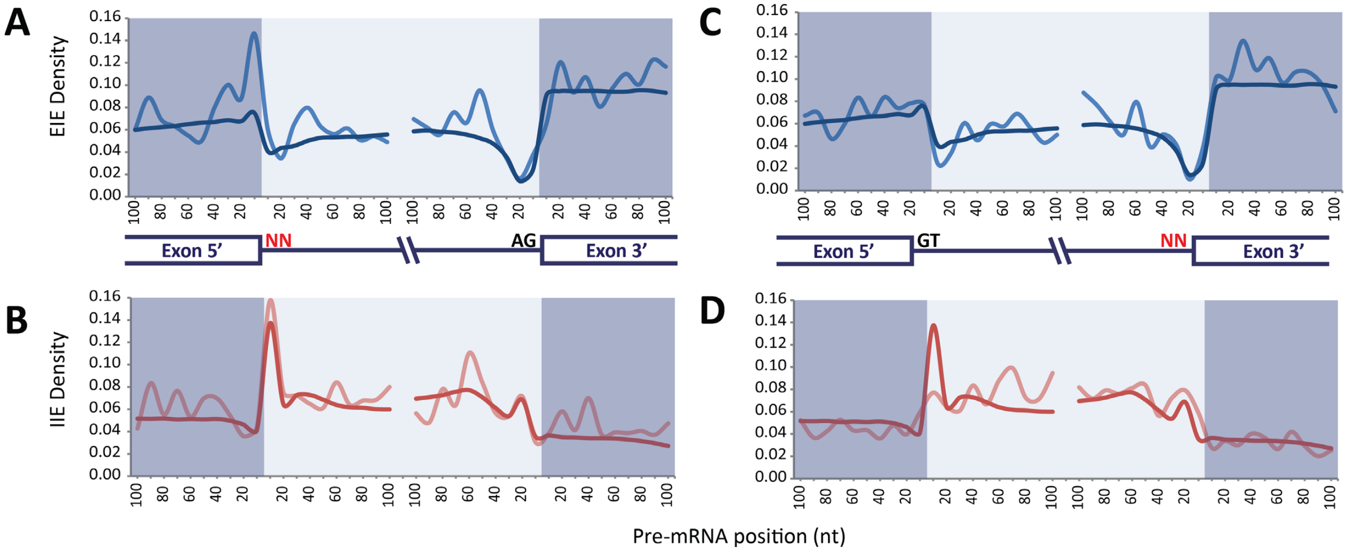

Figure 4. Non-canonical site junctions have distinctive distribution of EIE and IIE. (**A**) Positional density of EIE in canonical (blue) and 5′ non-canonical (pale blue) splice sites. (**B**) Positional density of IIE in canonical (red) and 5′ non-canonical (pale red) splice sites. (**C**) Positional density of EIE in canonical (blue) and 3′ non-canonical (pale blue) splice sites. (**D**) Positional density of IIE in canonical (red) and 3′ non-canonical (pale red) splice sites. The positional density is smoothed over a window of 10 bases.

some RNA-seq mappers, like TopHat (44), cannot align the reads to the non-canonical splice junction even though this alignment does not contain mismatches. Instead of this, TopHat aligns the reads with three mismatches near 5′-GT and AG-3′ splice sites. Moreover, this error is also present in the GENCODE annotation. The alignment of the cDNAs that we obtained by RT-PCR and Sanger sequencing corroborates the mRNA sequence and splicing of the non-canonical GA–AG intron (Figure 6A).

TopHat can only map reads to canonical splice junctions (GT–AG, GC–AG and AT–AC) (44–49). For this reason, when there are no putative canonical splice sites near to actual non-canonical splice junctions, TopHat is unable to align the RNA-seq reads. This is observed at ETV1 transcripts, where TopHat cannot align RNA-seq reads to a TT–AG non-canonical splice junction, so the reads cannot be assembled into one single transcript by Cufflinks (50) (Figure 6B).

The majority of non-canonical splice sites that do not have U2/U12-like sequences may be artifacts

The intron with non-canonical splice sites of XBP1 is the only mRNA intron that is known to be processed by non-spliceosome machinery (51), so it does not have U2/U12-like sequences. In the present analysis, we found 278 introns without U2/U12-like sequences around their splice sites (non-U2/U12) (Figure 1C). The fact that we detected the known non-canonical splice sites of XBP1 gene indicates that the methodology used for this work is sensitive to find splice junctions that are not processed by the spliceosome (Supplementary Figure S10).

Among the non-U2/U12 splice sites, we found 83 predicted splice sites that have their 5′ and 3′ ends near (≤10 nt) to a canonical splice junction (Supplementary Data). However, 65 of these predicted non-canonical splice sites are flanked by 5′ or 3′ exonic homopolymers of 5 nt or longer. The detection of these false non-canonical introns can be explained by misalignments to canonical splice junctions due to homopolymer associated indels.

Template switching during reverse transcription is also a potential source of false non-canonical splice junctions (52–55). When a RNA template contains direct repeats, RT can switch from one portion of the template to another generating a deletion in the cDNA. This phenomenon could generate false splicing-like events during cDNA library construction and it has been shown to be favored by the presence of direct repeats and secondary structures (52–55). We found 160 predicted non-canonical splice junctions that do not have U2/U12-like sequences and are not near to canonical splice sites (Supplementary Data). These predicted non-canonical splice junctions show longer direct repeats and higher %GC than canonical and non-canonical U2/U12-like splice junctions (Figure 7). Taken together, the data suggest that this group of predicted non-canonical splice junctions is highly enriched in template switching artifacts. We tested the veracity of the predicted non-U2/U12 splice junctions present in CCNG1, PSENEN and NAPA genes through RT-PCR and Sanger sequencing. The RT-PCR products that evidence the predicted non-U2/U12 splice junctions were obtained in highly variable proportion depending on which RT enzyme was used (Figure 8, Supplementary Figures S11 and S12). For example, we tested the reproducibility of the detection of CCNG1 non-U2/U12 splice junction by RT-PCR using MMLV or AMV RT enzymes. The CCNG1 non-U2/U12 splice junction was detected in MMLV RT-PCR products, but was not observed when AMV was used (Figure 8A and B). Similar results were obtained with the PSENEN and NAPA RT-PCRs (Supplementary Figures S11 and S12). Moreover, these predicted non-U2/U12 splice junctions are associated with

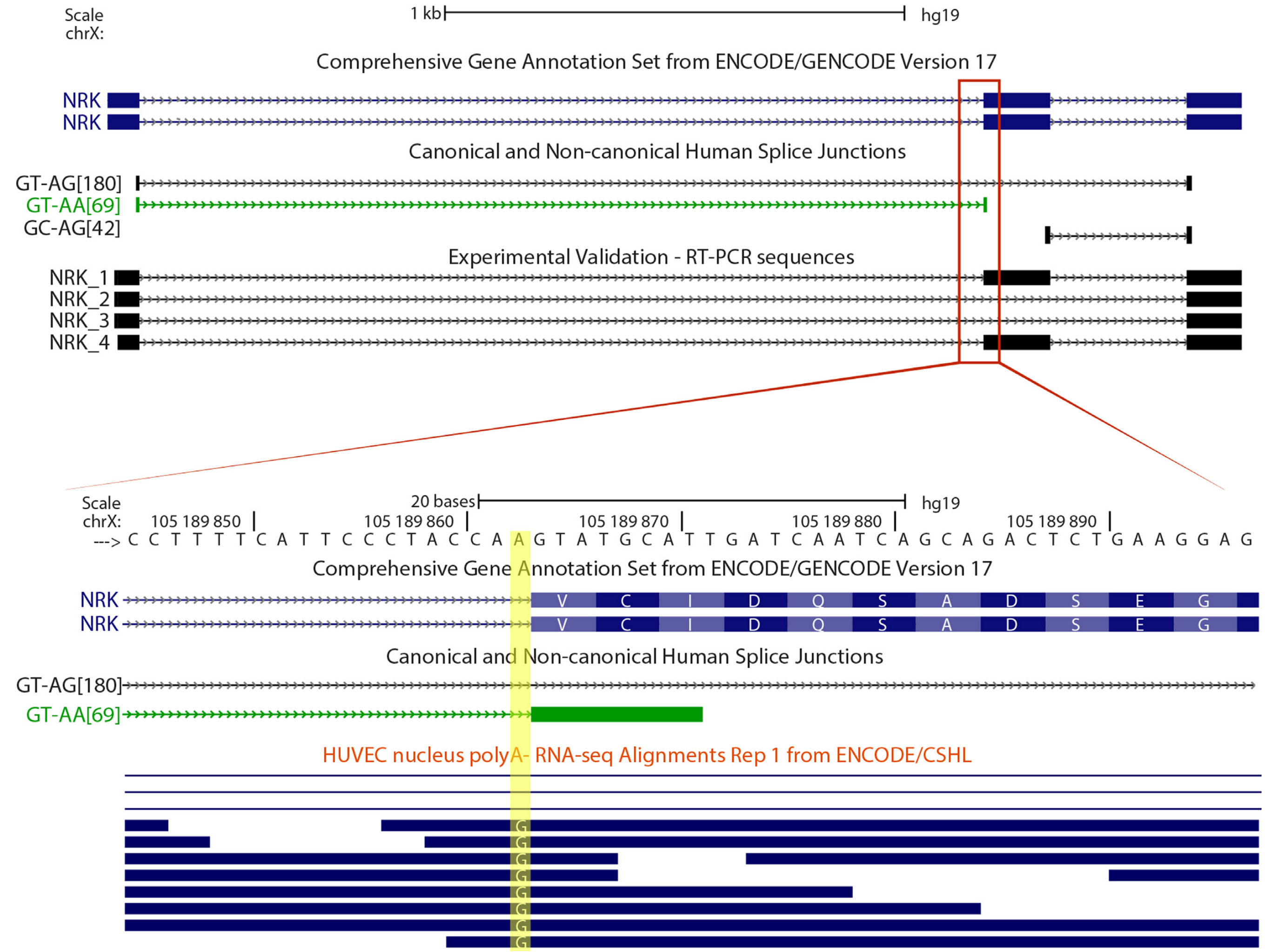
Scale
chrX:
1 kb
hg19
Comprehensive Gene Annotation Set from ENCODE/GENCODE Version 17
NRK
NRK
Canonical and Non-canonical Human Splice Junctions
GT-AG[180]
GT-AA[69]
GC-AG[42]
Experimental Validation - RT-PCR sequences
NRK_1
NRK_2
NRK_3
NRK_4
Scale
chrX:
20 bases
hg19
105 189 850
105 189 860
105 189 870
105 189 880
105 189 890
---> C C T T T T C A T T C C C T A C C A A G T A T G C A T T G A T C A A T C A G C A G A C T C T G A A G G A G
Comprehensive Gene Annotation Set from ENCODE/GENCODE Version 17
NRK
V C I D Q S A D S E G
NRK
V C I D Q S A D S E G
Canonical and Non-canonical Human Splice Junctions
GT-AG[180]
GT-AA[69]
HUVEC nucleus polyA- RNA-seq Alignments Rep 1 from ENCODE/CSHL
G
G
G
G
G
G
G
G

Figure 5. Editing-dependent splicing of non-canonical splice sites. The adenine of the AA-3′ non-canonical splice site that is highlighted in yellow shows a consistent A>G mismatch in poly(A)-minus RNA-seq data from HUVEC cell line. This reflects an A-to-I editing event in the AA-3′ non-canonical splice site, which likely allows the splicing of the GT–AA intron and the exon inclusion of a cassette exon. The exon skipping event is not annotated in GENCODE v17.

high %GC RNA secondary structures (Figure 8C, Supplementary Figures S11C and S12C).

DISCUSSION

We have done a comprehensive analysis of human non-canonical splice sites based on deep transcriptome sequencing data generated by RNA-seq. Our method was developed to avoid false non-canonical splice site sources, like alignment errors and polymorphisms. We found 184 U2/U12-like non-canonical splice sites and 51% of them are not annotated on GENCODE v.17. Our results represent a reliable catalog of non-canonical splice sites.

At least 28% of human U2/U12-like non-canonical splice sites are conserved in the mouse transcriptome. Even more, we found examples of U2/U12-like non-canonical splice sites that are highly conserved across vertebrates. The evolutionary conservation supports the idea that this group of splice sites is neither experimental artifact nor spliceosomal errors. Conversely, evolutionary conservation suggests a putative role of non-canonical splice sites in gene regulation. Furthermore, the human U2/U12-like non-canonical splice sites that are not conserved in mouse transcriptome may come from recent evolutionary events as the primate-specific non-canonical splice site that we found in ACTR10 (Figure 2B).

Our analysis shows that U2/U12-like non-canonical splice sites are highly involved in alternative splicing in comparison with canonical splice sites. Alternative splice sites are reported to be weaker than the constitutively processed splice sites and this weakening of splice sites has been proposed as a mechanism of alternative splicing through evolution (56,57). Thus, the weak splice signals of the U2/U12-like non-canonical splice sites could lead to alternative splice site activation. This is reflected on a splice junction in BANP gene in which only the vertebrates that have non-canonical splice sites have evidences of alternative 3′ splice site usage (Figure 2A).

Our data show that several alternative splicing events associated with U2/U12-like non-canonical splice sites

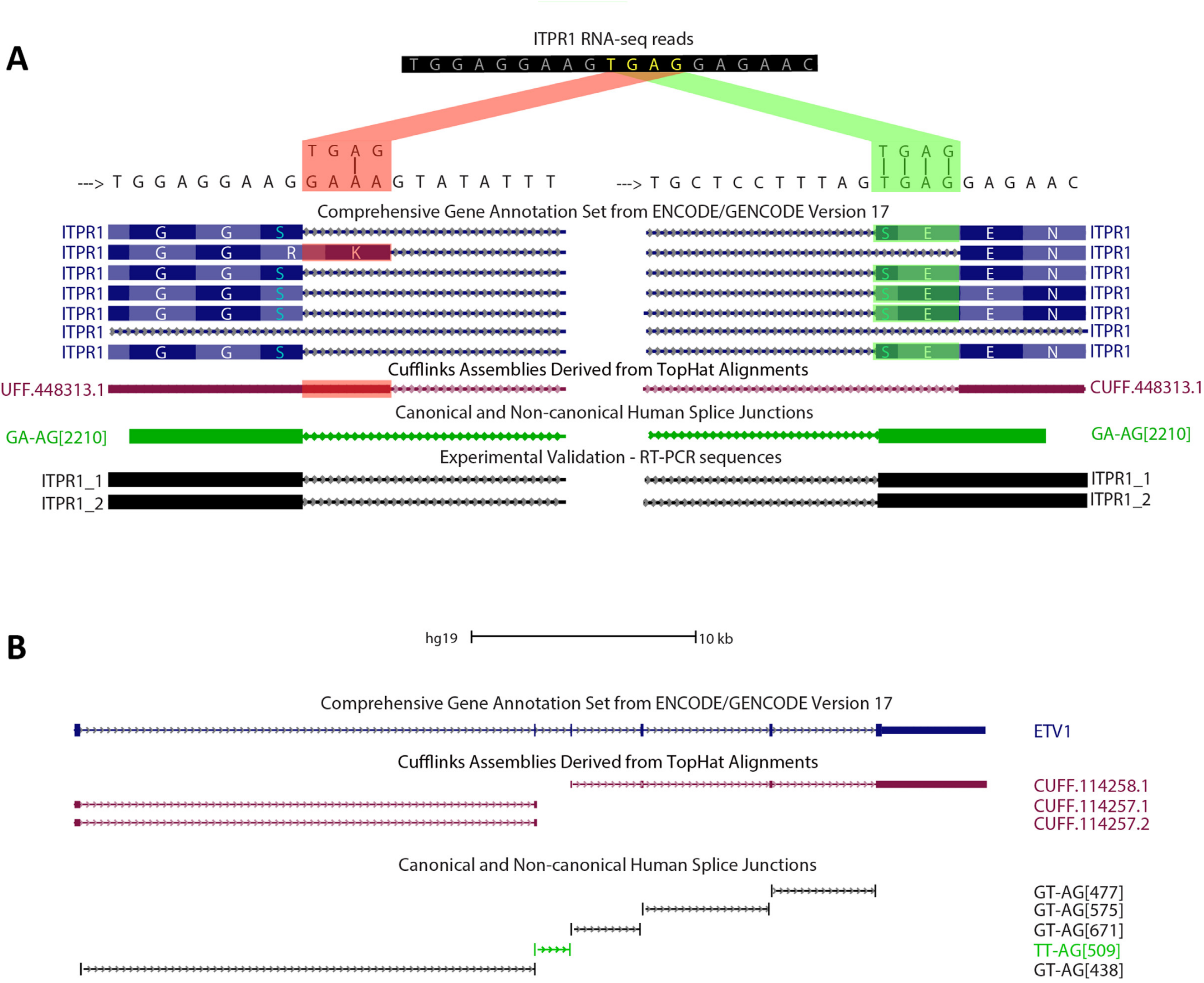
A
ITPR1 RNA-seq reads
T G G A G G A A G T G A G G A G A A C
T G A G
---> T G G A G G A A G G A A A G T A T A T T T
T G A G
---> T G C T C C T T T A G T G A G G A G A A C
Comprehensive Gene Annotation Set from ENCODE/GENCODE Version 17
ITPR1
G
S
R
K
E
N
Cufflinks Assemblies Derived from TopHat Alignments
CUFF.448313.1
Canonical and Non-canonical Human Splice Junctions
GA-AG[2210]
Experimental Validation - RT-PCR sequences
ITPR1_1
ITPR1_2
B
hg19
10 kb
Comprehensive Gene Annotation Set from ENCODE/GENCODE Version 17
ETV1
Cufflinks Assemblies Derived from TopHat Alignments
CUFF.114258.1
CUFF.114257.1
CUFF.114257.2
Canonical and Non-canonical Human Splice Junctions
GT-AG[477]
GT-AG[575]
GT-AG[671]
TT-AG[509]
GT-AG[438]

Figure 6. Non-canonical splice sites are prone to be misannotated. (**A**) UCSC Genome Browser image shows a splice site area of ITPR1 where RNA-seq reads can be aligned in two ways. The difference between the two alignments relies on a TGAG sequence (yellow letters) that can be aligned with three mismatches, evidencing a canonical GT–AG splice junction (red alignment) or without mismatches, but evidencing a non-canonical GA–AG splice junction (green alignment). A GENCODE isoform is based on the suboptimal alignment (highlighted in red). Assembled transcript of ITPR1 based on Tophat RNA-seq alignments is based on the suboptimal alignment (highlighted in red). RT-PCR coupled to Sanger sequencing probed that this transcript does not have any mismatches with the genome. (**B**) ETV1 gene has a constitutive TT–AG non-canonical splice junction that is annotated in GENCODE, but Cufflinks cannot assemble a continuous transcript for ETV1 due to TopHat's inability to align non-canonical splice junctions.

are tissue-specific. For instance, the CPSF3 non-canonical splice junction was only detected in testes. Others, like TSPYL2 non-canonical splice junction, show different levels of selection across human tissues. Moreover, the use of TSPYL2 non-canonical splice site generates a premature termination codon that disrupts a functional NAP-like motive and may induce non-sense mediated decay of TSPYL2 transcripts. This could be a regulatory mechanism of tissue-specific TSPYL2 expression, as has been shown for other mammalian genes (58–60).

SREs can recruit *trans*-acting splicing factors that modulate alternative splicing (61). We found an enrichment of SREs nearby non-canonical dinucleotides, the highest increase found in 5′ regions of 5′ non-canonical splice sites and 3′ regions of 3′ non-canonical splice sites. The SREs are shown to have a key role for alternative 5′/3′ selection (62–64), which are the most frequent type of alternative splicing events involving non-canonical splice sites. The weak splice signals of non-canonical splice sites and the high SRE density at their proximal regions indicate a high regulatory potential in the selection of these splice sites. In addition, this may explain the high percentage of non-canonical splice sites that are involved in alternative splicing events.

Most of the U2/U12-like splice sites have dinucleotides with only one mismatch with the canonical splice site dinucleotides. However, a single base mismatch at canonical U2 splice site dinucleotides induces dramatic functional impairments (3) and is frequently associated with mis-splicing diseases (65–69). Why these splice junctions can be processed even if they have non-canonical splice sites? It has

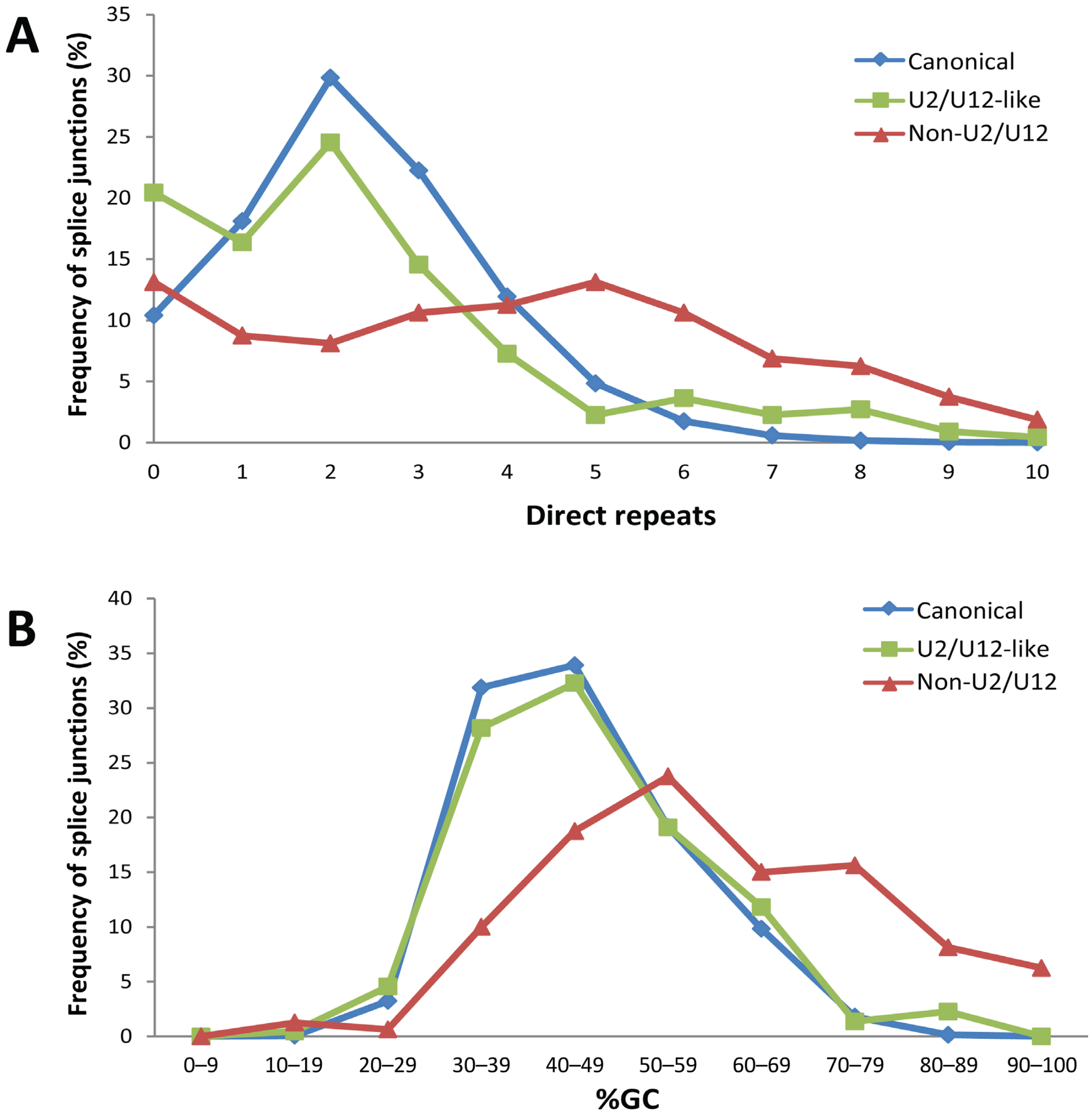

Figure 7. Non-U2/U12 splice junctions have a higher number of direct repeats and %GC content. (**A**) Frequency distribution of the number of direct repeats at canonical, U2/U12-like non-canonical and non-U2/U12 non-canonical splice junctions. (**B**) Frequency distribution of %GC content in canonical, U2/U12-like non-canonical and non-U2/U12 splice junctions.

been reported that two GT–AA and one AT–AG splice sites are converted into canonical splice sites by RNA A-to-I editing (41–43). We found five GT–AA and three AT–AG non-canonical splice sites that show evidence to be edited at their dinucleotides, and six of them are novel examples of this phenomenon (Table 2). Moreover, this RNA A-to-I editing enhances the selection of the edited non-canonical splice sites (Figure 5). However, we also found GT–AA and AT–AG non-canonical splice sites that do not show evidence of RNA A-to-I editing. The splicing of the remaining U2/U12-like splice sites cannot be explained by this mechanism.

Previous investigations have shown that U12 introns can be processed correctly by using combinations of terminal dinucleotides different from the known AT–AC and GT–AG. Analysis of splice sites in databases from different organisms (6) and point mutational analysis of U12-dependent splice sites (4) showed that a 5 ′A can splice to any 3′ nucleotide (AT–AN) although with a preference for 3′ AC dinucleotide. A 5 ′G can splice to 3 ′G or T nucleotide (GT–AG or GT–AT) with a preference for 3′ G nucleotide. This is consistent with our data as shown in Table 1. U12 spliceosome has a different mechanism of recognizing splice sites. It has been shown that the U11 and U12 snRNAs associate with each other before binding to the 5′ss and branch points, respectively (70). Since the U12 branch point plays a much bigger role in splice site recognition, this leads to a more relaxed constraint on the 3′ splice site dinucleotides. Burset *et al.* (13) proposed that non-canonical splice sites could function exclusively in association with near canonical splice sites that can efficiently recruit the splicing machinery, presenting a parasitic non-

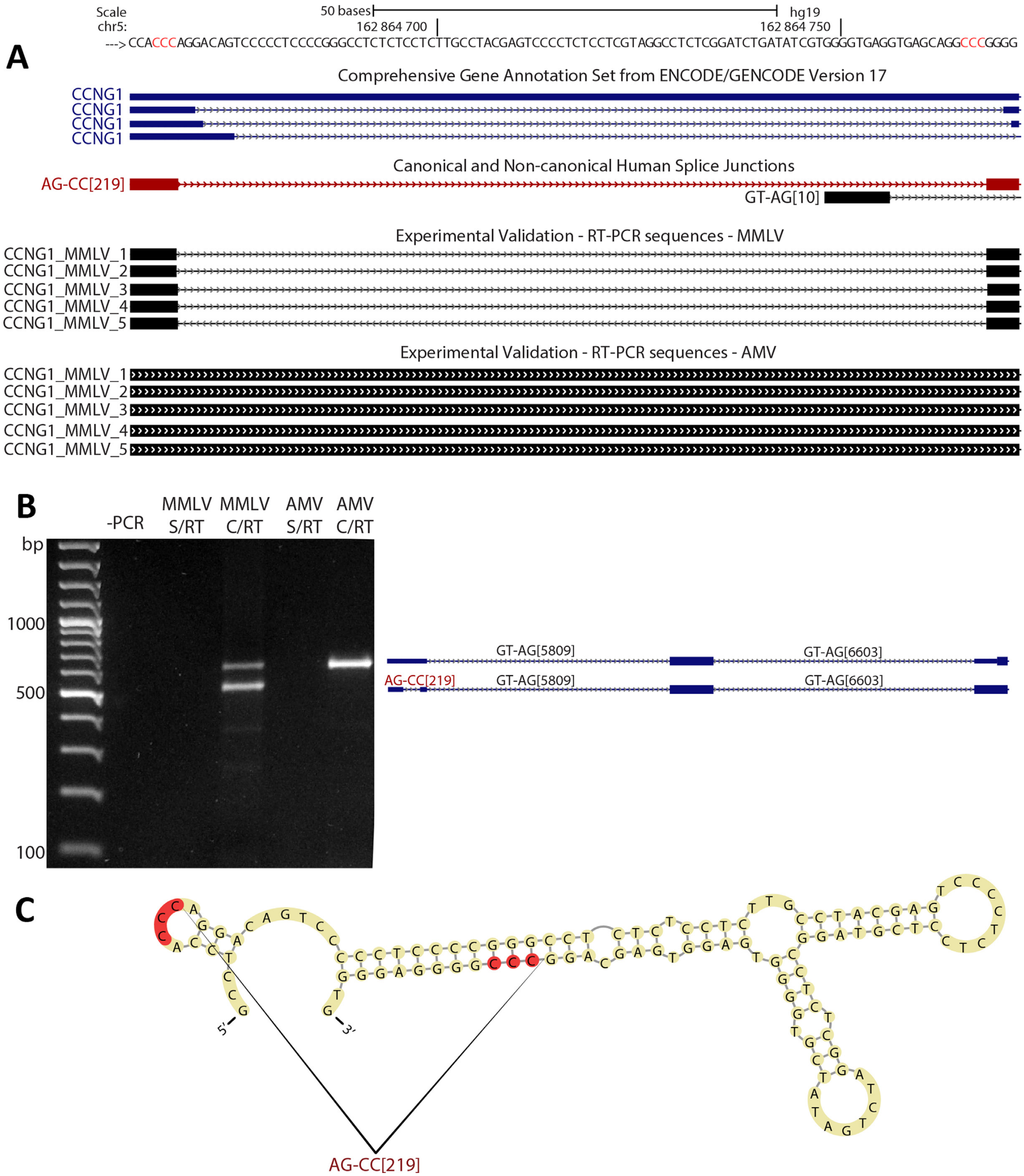
A
Scale
chr5:
50 bases
hg19
162 864 700
162 864 750
---> CCACCCAGGACAGTCCCCCTCCCCGGGCCTCTCTCCTCTTGCCTACGAGTCCCCTCTCCTCGTAGGCCTCTCGGATCTGATATCGTGGGGTGAGGTGAGCAGGCCCGGGG
Comprehensive Gene Annotation Set from ENCODE/GENCODE Version 17
CCNG1
CCNG1
CCNG1
CCNG1
Canonical and Non-canonical Human Splice Junctions
AG-CC[219]
GT-AG[10]
Experimental Validation - RT-PCR sequences - MMLV
CCNG1_MMLV_1
CCNG1_MMLV_2
CCNG1_MMLV_3
CCNG1_MMLV_4
CCNG1_MMLV_5
Experimental Validation - RT-PCR sequences - AMV
CCNG1_MMLV_1
CCNG1_MMLV_2
CCNG1_MMLV_3
CCNG1_MMLV_4
CCNG1_MMLV_5
B
-PCR
MMLV S/RT
MMLV C/RT
AMV S/RT
AMV C/RT
bp
1000
500
100
GT-AG[5809]
GT-AG[6603]
AG-CC[219]
GT-AG[5809]
GT-AG[6603]
C
5′
3′
AG-CC[219]

Figure 8. The non-U2/U12 splice junction of CCNG1 is a template switching artifact. (**A**) Our human splice junction annotation shows a non-U2/U12 splice junction (AG–CC[219]) present in the 5′ UTR of CCNG1 gene. Other non-U2/U12 splice junctions for this gene are annotated in GENCODE, but only the non-U2/U12 splice junction from our annotation (shown in red) was obtained by RT-PCR and Sanger sequencing. Red genomic letters indicate the 3-nt long direct repeat associated with this splice junction. (**B**) RT-PCR of CCNG1 transcripts using MMVL or AMV enzymes; alongside are represented the different products amplified. (**C**) *In silico* prediction of the secondary RNA structure associated with the CCNG1 non-U2/U12 splice junction. Directs repeats are highlighted in red.

canonical/canonical splice site relationship. This mechanism has been shown for the GA–AG non-canonical splice sites present on FGF1, where the 5′-GA non-canonical splice site recognition depends on 5′-GT canonical splice site situated 6 nt upstream and two intronic sequences downstream (9). However, more evidence is needed to know if all non-canonical splice sites depend on canonical splice sites to be recognized. Moreover, this mechanism cannot explain the constitutive non-canonical splice site processing that represents 24% of the non-canonical splice sites found by our analysis. Thus, more experimental evidence is necessary to understand the molecular mechanisms that contribute to non-canonical splice site selection.

Detection of non-canonical splice sites is often attributed to artifactual events (13,71–73). However, real non-canonical splice sites are frequently misannotated and underestimated by most used RNA-seq analysis tools (44–49). The presence of non-canonical splice sites at human transcripts generates false canonical splice site detection or disruption of assembled transcripts, evidencing that transcript annotation pipelines and RNA-seq tools have a deficient methodology to do *ab initio* detection of non-canonical splice sites. Thus, our annotation of non-canonical splice sites will lead to an improvement of the human transcriptome annotation.

In the present analysis, extensive methods were designed to avoid false non-canonical splice site detection. However, the non-U2/U12 splice sites represent a group of splice sites that could be highly affected by sequencing errors and reverse transcription artifacts during cDNA library construction. The Illumina sequencing platform, used to generate the RNA-seq data, is considered robust against indel errors; however, within homopolymers, the indel error rate is dramatically higher (74). We found 64 non-U2/U12 splice sites that are flanked by 5′ or 3′ exonic homopolymers. These non-U2/U12 splice sites are ≤10 nt from canonical splice sites. Thus, non-U2/U12 splice sites are more likely to be homopolymer-associated indels than U2/U12-like splice sites.

False non-canonical splice sites can be generated by RT template switching, a phenomenon that is enhanced by secondary RNA structures and the presence of direct repeats at the splice junctions (52–55). The group of non-U2/U12 splice sites identified herein has longer direct repeats and higher %GC than canonical and non-canonical U2/U12-like introns, suggesting an enrichment of reverse transcription artifact in non-U2/U12 introns. RT-PCRs designed to amplify non-U2/U12 splice junctions present on CCNG1, PSENEN and NAPA genes show different results depending on the RTs used. These splice junctions are associated with high %GC introns and RNA secondary structures, suggesting that at least these three non-canonical splice sites are artifacts. These results indicate that a RT-independent sequencing method of RNA is needed to make a confident catalog of non-U2/U12 splices sites.

In conclusion, here we report the landscape of non-canonical splice sites present at the human transcriptome. We identified a high confidence group of 184 U2/U12-like non-canonical splice sites that will improve the human transcriptome annotation. Non-canonical splice site selection is a rare event in human transcriptome associated with distinctive features. Almost all U2/U12-like non-canonical splice sites have only one non-canonical dinucleotide, which is often associated with alternative canonical dinucleotide. Additionally, the high SRE density of exon–intron boundaries that have U2/U12-like non-canonical splice sites gives an appropriated context to process these weak splice sites in a regulated fashion.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGMENTS

We want to thank Dr María Estela Andrés and Alejandro Montenegro for their valuable comments on the manuscript.

FUNDING

Millennium Science Initiative [P10-063/F]; FONDECYT [1110392]; CONICYT fellowship [to R.M.]. Funding for open access charge: FONDECYT [1110392].

Conflict of interest statement. None declared.

REFERENCES

1. Hoskins,A.A. and Moore,M.J. (2012) The spliceosome: a flexible, reversible macromolecular machine. *Trends Biochem. Sci.*, **37**, 179–188.
2. Will,C.L. and Luhrmann,R. (2011) Spliceosome structure and function. *Cold Spring Harb. Perspect. Biol.*, **3**.
3. Aebi,M., Hornig,H., Padgett,R.A., Reiser,J. and Weissmann,C. (1986) Sequence requirements for splicing of higher eukaryotic nuclear pre-mRNA. *Cell*, **47**, 555–565.
4. Dietrich,R.C., Fuller,J.D. and Padgett,R.A. (2005) A mutational analysis of U12-dependent splice site dinucleotides. *RNA*, **11**, 1430–1440.
5. Lamond,A.I., Konarska,M.M. and Sharp,P.A. (1987) A mutational analysis of spliceosome assembly: evidence for splice site collaboration during spliceosome formation. *Genes Dev.*, **1**, 532–543.
6. Sheth,N., Roca,X., Hastings,M.L., Roeder,T., Krainer,A.R. and Sachidanandam,R. (2006) Comprehensive splice-site analysis using comparative genomics. *Nucleic Acids Res.*, **34**, 3955–3967.
7. Patel,A.A. and Steitz,J.A. (2003) Splicing double: insights from the second spliceosome. *Nat. Rev. Mol. Cell Biol.*, **4**, 960–970.
8. Twigg,S.R., Burns,H.D., Oldridge,M., Heath,J.K. and Wilkie,A.O. (1998) Conserved use of a non-canonical 5’ splice site (/GA) in alternative splicing by fibroblast growth factor receptors 1, 2 and 3. *Hum. Mol. Genet.*, **7**, 685–691.
9. Brackenridge,S., Wilkie,A.O. and Screaton,G.R. (2003) Efficient use of a ‘dead-end’ GA 5’ splice site in the human fibroblast growth factor receptor genes. *EMBO J.*, **22**, 1620–1631.
10. Quan,F. and Forte,M.A. (1990) Two forms of Drosophila melanogaster Gs alpha are produced by alternate splicing involving an unusual splice site. *Mol. Cell. Biol.*, **10**, 910–917.
11. Pollard,A.J., Krainer,A.R., Robson,S.C. and Europe-Finner,G.N. (2002) Alternative splicing of the adenylyl cyclase stimulatory G-protein G alpha(s) is regulated by SF2/ASF and heterogeneous nuclear ribonucleoprotein A1 (hnRNPA1) and involves the use of an unusual TG 3’-splice site. *J. Biol. Chem.*, **277**, 15241–15251.
12. Szafranski,K., Schindler,S., Taudien,S., Hiller,M., Huse,K., Jahn,N., Schreiber,S., Backofen,R. and Platzer,M. (2007) Violating the splicing rules: TG dinucleotides function as alternative 3’ splice sites in U2-dependent introns. *Genome Biol.*, **8**, R154.
13. Burset,M., Seledtsov,I.A. and Solovyev,V.V. (2000) Analysis of canonical and non-canonical splice sites in mammalian genomes. *Nucleic Acids Res.*, **28**, 4364–4375.

14. Wang,Z., Gerstein,M. and Snyder,M. (2009) RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.*, **10**, 57–63.
15. Pan,Q., Shai,O., Lee,L.J., Frey,B.J. and Blencowe,B.J. (2008) Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nat. Genet.*, **40**, 1413–1415.
16. Sultan,M., Schulz,M.H., Richard,H., Magen,A., Klingenhoff,A., Scherf,M., Seifert,M., Borodina,T., Soldatov,A., Parkhomchuk,D. *et al.* (2008) A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome. *Science*, **321**, 956–960.
17. Wang,E.T., Sandberg,R., Luo,S., Khrebtukova,I., Zhang,L., Mayr,C., Kingsmore,S.F., Schroth,G.P. and Burge,C.B. (2008) Alternative isoform regulation in human tissue transcriptomes. *Nature*, **456**, 470–476.
18. Djebali,S., Davis,C.A., 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.
19. Rozowsky,J., Abyzov,A., Wang,J., Alves,P., Raha,D., Harmanci,A., Leng,J., Bjornson,R., Kong,Y., Kitabayashi,N. *et al.* (2011) AlleleSeq: analysis of allele-specific expression and binding in a network framework. *Mol. Syst. Biol.*, **7**, 522.
20. Wang,K., Singh,D., Zeng,Z., Coleman,S.J., Huang,Y., Savich,G.L., He,X., Mieczkowski,P., Grimm,S.A., Perou,C.M. *et al.* (2010) MapSplice: accurate mapping of RNA-seq reads for splice junction discovery. *Nucleic Acids Res.*, **38**, e178.
21. Kent,W.J. (2002) BLAT—the BLAST-like alignment tool. *Genome Res.*, **12**, 656–664.
22. Kapitonov,V.V. and Jurka,J. (2008) A universal classification of eukaryotic transposable elements implemented in Repbase. *Nat. Rev. Genet.*, **9**, 411–412.
23. Morgulis,A., Gertz,E.M., Schaffer,A.A. and Agarwala,R. (2006) A fast and symmetric DUST implementation to mask low-complexity DNA sequences. *J. Comput. Biol.*, **13**, 1028–1040.
24. Rhead,B., Karolchik,D., Kuhn,R.M., Hinrichs,A.S., Zweig,A.S., Fujita,P.A., Diekhans,M., Smith,K.E., Rosenbloom,K.R., Raney,B.J. *et al.* (2010) The UCSC Genome Browser database: update 2010. *Nucleic Acids Res.*, **38**, D613–D619.
25. Langmead,B., Trapnell,C., Pop,M. and Salzberg,S.L. (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.*, **10**, R25.
26. Fairbrother,W.G., Yeh,R.F., Sharp,P.A. and Burge,C.B. (2002) Predictive identification of exonic splicing enhancers in human genes. *Science*, **297**, 1007–1013.
27. Wang,Z., Rolish,M.E., Yeo,G., Tung,V., Mawson,M. and Burge,C.B. (2004) Systematic identification and analysis of exonic splicing silencers. *Cell*, **119**, 831–845.
28. Wang,Y., Ma,M., Xiao,X. and Wang,Z. (2012) Intronic splicing enhancers, cognate splicing factors and context-dependent regulation rules. *Nat. Struct. Mol. Biol.*, **19**, 1044–1052.
29. Wang,Y., Xiao,X., Zhang,J., Choudhury,R., Robertson,A., Li,K., Ma,M., Burge,C.B. and Wang,Z. (2013) A complex network of factors with overlapping affinities represses splicing through intronic elements. *Nat. Struct. Mol. Biol.*, **20**, 36–45.
30. Zhang,C., Li,W.H., Krainer,A.R. and Zhang,M.Q. (2008) RNA landscape of evolution for optimal exon and intron discrimination. *Proc. Natl. Acad. Sci. U.S.A.*, **105**, 5797–5802.
31. Korbie,D.J. and Mattick,J.S. (2008) Touchdown PCR for increased specificity and sensitivity in PCR amplification. *Nat. Protoc.*, **3**, 1452–1456.
32. Gruber,A.R., Lorenz,R., Bernhart,S.H., Neubock,R. and Hofacker,I.L. (2008) The Vienna RNA websuite. *Nucleic Acids Res.*, **36**, W70–W74.
33. Byun,Y. and Han,K. (2009) PseudoViewer3: generating planar drawings of large-scale RNA structures with pseudoknots. *Bioinformatics*, **25**, 1435–1437.
34. Sherry,S.T., Ward,M.H., Kholodov,M., Baker,J., Phan,L., Smigielski,E.M. and Sirotkin,K. (2001) dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.*, **29**, 308–311.
35. Engstrom,P.G., Steijger,T., Sipos,B., Grant,G.R., Kahles,A., Consortium,R., Alioto,T., Behr,J., Bertone,P., Bohnert,R. *et al.* (2013) Systematic evaluation of spliced alignment programs for RNA-seq data. *Nat. Methods*, **10**, 1185–1191.
36. Li,H. and Homer,N. (2010) A survey of sequence alignment algorithms for next-generation sequencing. *Brief. Bioinform.*, **11**, 473–483.
37. Ameur,A., Wetterbom,A., Feuk,L. and Gyllensten,U. (2010) Global and unbiased detection of splice junctions from RNA-seq data. *Genome Biol.*, **11**, R34.
38. The UniProt Consortium. (2013) Update on activities at the Universal Protein Resource (UniProt) in 2013. *Nucleic Acids Res.*, **41**, D43–D47.
39. Barash,Y., Calarco,J.A., Gao,W., Pan,Q., Wang,X., Shai,O., Blencowe,B.J. and Frey,B.J. (2010) Deciphering the splicing code. *Nature*, **465**, 53–59.
40. Wang,Y. and Wang,Z. (2014) Systematical identification of splicing regulatory cis-elements and cognate trans-factors. *Methods*, **65**, 350–358.
41. Rueter,S.M., Dawson,T.R. and Emeson,R.B. (1999) Regulation of alternative splicing by RNA editing. *Nature*, **399**, 75–80.
42. Lev-Maor,G., Sorek,R., Levanon,E.Y., Paz,N., Eisenberg,E. and Ast,G. (2007) RNA-editing-mediated exon evolution. *Genome Biol.*, **8**, R29.
43. Athanasiadis,A., Rich,A. and Maas,S. (2004) Widespread A-to-I RNA editing of Alu-containing mRNAs in the human transcriptome. *PLoS Biol.*, **2**, e391.
44. Trapnell,C., Pachter,L. and Salzberg,S.L. (2009) TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics*, **25**, 1105–1111.
45. Au,K.F., Jiang,H., Lin,L., Xing,Y. and Wong,W.H. (2010) Detection of splice junctions from paired-end RNA-seq data by SpliceMap. *Nucleic Acids Res.*, **38**, 4570–4578.
46. Lou,S.K., Ni,B., Lo,L.Y., Tsui,S.K., Chan,T.F. and Leung,K.S. (2011) ABMapper: a suffix array-based tool for multi-location searching and splice-junction mapping. *Bioinformatics*, **27**, 421–422.
47. Wu,J., Anczukow,O., Krainer,A.R., Zhang,M.Q. and Zhang,C. (2013) OLego: fast and sensitive mapping of spliced mRNA-Seq reads using small seeds. *Nucleic Acids Res.*, **41**, 5149–5163.
48. Huang,S., Zhang,J., Li,R., Zhang,W., He,Z., Lam,T.W., Peng,Z. and Yiu,S.M. (2011) SOAPsplice: Genome-Wide ab initio Detection of Splice Junctions from RNA-Seq Data. *Front. Genet.*, **2**, 46.
49. Denoeud,F., Aury,J.M., Da Silva,C., Noel,B., Rogier,O., Delledonne,M., Morgante,M., Valle,G., Wincker,P., Scarpelli,C. *et al.* (2008) Annotating genomes with massive-scale RNA sequencing. *Genome Biol.*, **9**, R175.
50. Trapnell,C., Roberts,A., Goff,L., Pertea,G., Kim,D., Kelley,D.R., Pimentel,H., Salzberg,S.L., Rinn,J.L. and Pachter,L. (2012) Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat. Protoc.*, **7**, 562–578.
51. Yoshida,H., Matsui,T., Yamamoto,A., Okada,T. and Mori,K. (2001) XBP1 mRNA is induced by ATF6 and spliced by IRE1 in response to ER stress to produce a highly active transcription factor. *Cell*, **107**, 881–891.
52. Cocquet,J., Chong,A., Zhang,G. and Veitia,R.A. (2006) Reverse transcriptase template switching and false alternative transcripts. *Genomics*, **88**, 127–131.
53. Mader,R.M., Schmidt,W.M., Sedivy,R., Rizovski,B., Braun,J., Kalipciyan,M., Exner,M., Steger,G.G. and Mueller,M.W. (2001) Reverse transcriptase template switching during reverse transcriptase-polymerase chain reaction: artificial generation of deletions in ribonucleotide reductase mRNA. *J. Lab. Clin. Med.*, **137**, 422–428.
54. Roy,S.W. and Irimia,M. (2008) When good transcripts go bad: artifactual RT-PCR 'splicing' and genome analysis. *Bioessays*, **30**, 601–605.
55. Houseley,J. and Tollervey,D. (2010) Apparent non-canonical trans-splicing is generated by reverse transcriptase in vitro. *PLoS One*, **5**, e12271.
56. Merkin,J., Russell,C., Chen,P. and Burge,C.B. (2012) Evolutionary dynamics of gene and isoform regulation in Mammalian tissues. *Science*, **338**, 1593–1599.
57. Ast,G. (2004) How did alternative splicing evolve? *Nat. Rev. Genet.*, **5**, 773–782.
58. Lewis,B.P., Green,R.E. and Brenner,S.E. (2003) Evidence for the widespread coupling of alternative splicing and nonsense-mediated mRNA decay in humans. *Proc. Natl. Acad. Sci. U.S.A.*, **100**, 189–192.

59. Baek,D. and Green,P. (2005) Sequence conservation, relative isoform frequencies, and nonsense-mediated decay in evolutionarily conserved alternative splicing. *Proc. Natl. Acad. Sci. U.S.A.*, **102**, 12813–12818.
60. Saltzman,A.L., Kim,Y.K., Pan,Q., Fagnani,M.M., Maquat,L.E. and Blencowe,B.J. (2008) Regulation of multiple core spliceosomal proteins by alternative splicing-coupled nonsense-mediated mRNA decay. *Mol. Cell. Biol.*, **28**, 4320–4330.
61. Matlin,A.J., Clark,F. and Smith,C.W. (2005) Understanding alternative splicing: towards a cellular code. *Nat. Rev. Mol. Cell Biol.*, **6**, 386–398.
62. Wang,Z., Xiao,X., Van Nostrand,E. and Burge,C.B. (2006) General and specific functions of exonic splicing silencers in splicing control. *Mol. Cell*, **23**, 61–70.
63. Yu,Y., Maroney,P.A., Denker,J.A., Zhang,X.H., Dybkov,O., Luhrmann,R., Jankowsky,E., Chasin,L.A. and Nilsen,T.W. (2008) Dynamic regulation of alternative splicing by silencers that modulate 5' splice site competition. *Cell*, **135**, 1224–1236.
64. Roca,X., Krainer,A.R. and Eperon,I.C. (2013) Pick one, but be quick: 5' splice sites and the problems of too many choices. *Genes Dev.*, **27**, 129–144.
65. D'Souza-Li,L., Canaff,L., Janicic,N., Cole,D.E. and Hendy,G.N. (2001) An acceptor splice site mutation in the calcium-sensing receptor (CASR) gene in familial hypocalciuric hypercalcemia and neonatal severe hyperparathyroidism. *Hum. Mutat.*, **18**, 411–421.
66. Lee,H.H. and Chang,S.F. (2001) Multiple transcripts of the CYP21 gene are generated by the mutation of the splicing donor site in intron 2 from GT to AT in 21-hydroxylase deficiency. *J. Endocrinol.*, **171**, 397–402.
67. D'Andrea,G., Di Perna,P., Brancaccio,V., Faioni,E.M., Castaman,G., Cibelli,G., Di Minno,G. and Margaglione,M. (2003) A novel G-to-A mutation in the intron-N of the protein S gene leading to abnormal RNA splicing in a patient with protein S deficiency. *Haematologica*, **88**, 459–464.
68. Scheckenbach,K., Balz,V., Wagenmann,M. and Hoffmann,T.K. (2008) An intronic alteration of the fibroblast growth factor 10 gene causing ALSG-(aplasia of lacrimal and salivary glands) syndrome. *BMC Med. Genet.*, **9**, 114.
69. Baralle,D. and Baralle,M. (2005) Splicing in action: assessing disease causing sequence changes. *J. Med. Genet.*, **42**, 737–748.
70. Frilander,M.J. and Steitz,J.A. (1999) Initial recognition of U12-dependent introns requires both U11/5' splice-site and U12/branchpoint interactions. *Genes Dev.*, **13**, 851–863.
71. Harrow,J., Denoeud,F., Frankish,A., Reymond,A., Chen,C.K., Chrast,J., Lagarde,J., Gilbert,J.G., Storey,R., Swarbreck,D. *et al.* (2006) GENCODE: producing a reference annotation for ENCODE. *Genome Biol.*, **7**(Suppl. 1), S4.
72. Sorek,R. and Safer,H.M. (2003) A novel algorithm for computational identification of contaminated EST libraries. *Nucleic Acids Res.*, **31**, 1067–1074.
73. Bonizzoni,P., Rizzi,R. and Pesole,G. (2005) ASPIC: a novel method to predict the exon-intron structure of a gene that is optimally compatible to a set of transcript sequences. *BMC Bioinformatics*, **6**, 244.
74. Minoche,A.E., Dohm,J.C. and Himmelbauer,H. (2011) Evaluation of genomic high-throughput sequencing data generated on Illumina HiSeq and genome analyzer systems. *Genome Biol.*, **12**, R112.