Human Molecular Genetics, 2016, Vol. 25, No. 22 4962–4982

doi: 10.1093/hmg/ddw321
Advance Access Publication Date: 20 September 2016
Original Article

ORIGINAL ARTICLE

Widespread splicing of repetitive element loci into coding regions of gene transcripts

Miranda M. Darby[1], Jeffrey T. Leek[2], Ben Langmead[3], Robert H. Yolken[1] and Sarven Sabunciyan[1,*]

[1]Department of Pediatrics, Johns Hopkins University, Baltimore, USA, [2]Department of Biostatistics, The Johns Hopkins University Bloomberg School of Public Health, Baltimore, USA and [3]Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Center for Computational Biology and Department of Computer Science, Johns Hopkins University, Baltimore, MD, USA

*To whom correspondence should be addressed at: Sarven Sabunciyan, Johns Hopkins University, Department of Pediatrics, 600 N. Wolfe Street, Blalock 1147, Baltimore, MD 21287, USA. Tel: 410-614-3918; Fax: 410-955-3723; Email: ssabunc1@jhmi.edu

Abstract

We performed a thorough characterization of expressed repetitive element loci (RE) in the human orbitofrontal cortex (OFC) using directional RNA sequencing data. Considering only sequencing reads that map uniquely onto the human genome, we discovered that the overwhelming majority of intronic and exonic RE are expressed in the same orientation as the gene in which they reside. Our mapping approach enabled the identification of novel differentially expressed RE transcripts between the OFC and peripheral blood lymphocytes. Further analysis revealed that RE are extensively spliced into coding regions of gene transcripts yielding thousands of novel mRNA variants with altered coding potential. Lower frequency splicing of RE into untranslated regions of gene transcripts was also observed. The same pattern of RE splicing in the brain was also detected for Drosophila, zebrafish, mouse, rat, dog and rabbit. RE splicing occurs largely at canonical GT-AG splice junctions with LINE and SINE elements forming the most RE splice junctions in the human OFC. This type of splicing usually gives rise to a minor splice variant of the endogenous gene and *in silico* analysis suggests that RE splicing has the potential to introduce novel open reading frames. Reanalysis of previously published sequencing data performed in the mouse cerebellum revealed that thousands of RE splice variants are associated with translating ribosomes. Our results demonstrate that RE expression is more complex than previously envisioned and raise the possibility that RE splicing might generate functional protein isoforms.

Introduction

The genomes of most eukaryotes contain large numbers of repetitive element loci (RE), which are primarily remnants of ancient transposition events (1). These elements are rendered inactive by mechanisms such as miRNA silencing and hypermethylation (2) and accumulate mutations that hinder their ability to transpose. Although there are more than 5 million annotated repetitive elements in humans, which make up nearly half of the genome, only very few of these elements are capable of transposing (3). These transposition competent elements have caused a number of cases of diseases (4). RE that are incorporated into the germline are inherited between generations and persist in the genome. The conventional thinking was that transposition-defective repetitive elements did not have a function, which made their persistence in the genome perplexing. New research has discovered that RE give rise to nearly one third of alternative promoter sites in the human genome and they facilitate the formation of gene networks (5,6). A small number of repetitive elements are reported to be exonized (7,8)

Received: April 8, 2016. **Revised:** August 22, 2016. **Accepted:** September 14, 2016

and certain repetitive element proteins are recruited to perform cellular functions (9).

Despite these findings, RE are routinely excluded from genomic, epigenetic and RNA expression studies. Our previous work demonstrated that RE are abundantly expressed in the human brain cortex (10), with almost 10% of the RNA sequencing reads originating from an RE. However, the extent to which each locus is transcribed is unknown. Therefore, we specifically investigated expression from individual RE in the cortex by performing strand specific RNA sequencing in 59 human orbitofrontal cortex samples (OFC). Most RE expression we detect is likely read through from gene and other promoters. However, we also find widespread splicing of RE into gene transcripts and have identified transcripts from 10238 different genes in the human OFC that have at least one RE splice junction. RE splice junctions occur predominantly in coding regions of gene transcripts and give rise to novel splice variants with altered coding potential. In addition, we find a total of 11041 different transcripts containing RE splice junctions in three different cerebellum cell types to associate with translating ribosomes. Our results demonstrate that RE expression is more complex than previously envisioned and raise the possibility that RE splicing may generate novel protein isoforms by extending the open reading frame of endogenous gene transcripts.

Results

RNA sequencing and RE expression analysis pipeline

We performed RNA-seq on the postmortem orbitofrontal cortex (OFC) samples from the Stanley Neuropathology Consortium Collection (11) (Supplementary Material, Table S1). This collection contains samples from 15 bipolar disorder, 15 schizophrenia, 15 depression and 15 control subjects. Previous studies on these and other postmortem brain samples have demonstrated that only subtle changes in transcription are associated with psychiatric disease (12,13). Thus, we decided to include results from samples obtained from individuals with psychiatric disease in our analysis. The raw sequencing data are available for download at http://sncid.stanleyresearch.org/; date last accessed September 26, 2016.

Strand-specific RNA-seq libraries were constructed and 100 base pair single end sequencing was performed (Supplementary Material, Table S2). To estimate expression from RE we exploited the fact that RE sequences drift over time. Since the vast majority of RE contain stretches of sequence that are unique in the human genome, we can measure the expression at specific loci based on RNA sequencing coverage at these unique regions. We aligned sequencing reads to the human genome and only considered reads with a mapping quality score of 40 or higher, corresponding to a one in 10,000 chance that the sequenced RNA originated from a different location (14). We estimate that more than 96% of RE annotated in the human genome contain at least one unique region that can be mapped unambiguously by RNAseq (see Uniqueness Estimates section in Methods).

To determine the capacity of our mapping approach to correctly map transcription of RE and detect differential expression, we performed unsupervised clustering on our cohort, which consists of both sexes, based on the expression of RE located on the Y chromosome. Loci that had low read counts varied greatly between samples of the same sex, thus we decided to only examine Y chromosome loci in which the sum of the reads that uniquely mapped to the genome in the 59 samples was at least 100. The 648 loci that exceeded this threshold, with 35 loci expressing from both strands, were used for the cluster analysis (for read counts see Supplementary Material, Table S3). In this analysis (Fig. 1) , males and females segregated into clear clusters, with females showing almost no expression of Y chromosome RE, indicating that RE are correctly mapped using our method.

An in depth analysis of our data revealed that only 4 loci had an average read count of more than one read in the female samples, with one locus having a read count of more than five. Reads align to this locus (chrY:21153221-21153519) in females due to a known annotation error. Reads appear to map uniquely onto the Y chromosome because the CD24 gene is missing on chromosome 6 in the hg19 genome assembly (15). In the male samples, every RE on the Y chromosome had an average count of more than one read with greater than sixty percent of loci having an average read count of five or more reads. These results suggest that errors in our analysis strategy can arise from noise that is inherent in low abundance transcripts and annotation errors. We remedied the problem of noise by introducing a minimum threshold level in all subsequent analyses. The issue of annotation errors, which also impact gene expression estimates, cannot be easily corrected. Fortunately, our analysis on the Y chromosome suggests that this is a minor problem with only one out of the 648 expressed loci being incorrectly annotated.

Expressed RE in OFC

We chose a threshold of more than 20 reads in more than three quarters of the samples to define expressed RE in the OFC. We selected this threshold based on the Y chromosome analysis in which correctly annotated RE had less than five reads in females. Note that the vast majority of the RE included in the Y chromosome cluster analysis (Fig. 1) fail to meet this threshold for expression. Although increasing this cutoff by four-fold may be too stringent, it provides additional confidence that loci meeting this criterion are expressed. Using this cutoff, we detected expression from 31,351 brain expressed RE (beRE) in the OFC. More than 70% of the beRE are located in genic regions and expressed in the same direction as the corresponding gene irrespective of the annotated orientation of the RE (Table 1) . The relative lack of exonic and intronic beRE that are expressed antisense to the corresponding gene suggests that most beRE in genes are likely transcribed by the promoter of the gene in which they reside. However, approximately 29% of beRE transcripts are located within intergenic regions and may be independently transcribed. To determine whether intergenic beRE reside in annotated non-coding transcripts, we compared beRE to all large intergenic non-coding RNAs (lincRNAs), and transcripts of uncertain coding potential (TUCPs) from the Human Body Map (16), as well as small noncoding RNAs including known and predicted snoRNAs, scaRNAs and microRNAs from snoRNABase (17) and miRBase (18). Only 368 (3.5%) of the intergenic beRE are located in lincRNAs or TUCPs and only two beRE overlap microRNAs.

We then categorized the expressed RE by class and found that approximately 40% are SINE elements, 23% are LINE elements, 14% are LTR elements, 10% are simple repeats, 6% are DNA transposons and 6% are low complexity regions (Supplementary Material, Table S4; this table also includes information on RE subfamilies). This pattern of expression was consistent between exons, introns and intergenic regions. Given that SINE elements make up 34% and LINE elements make up 28% of all annotated RE in the human genome (Supplementary

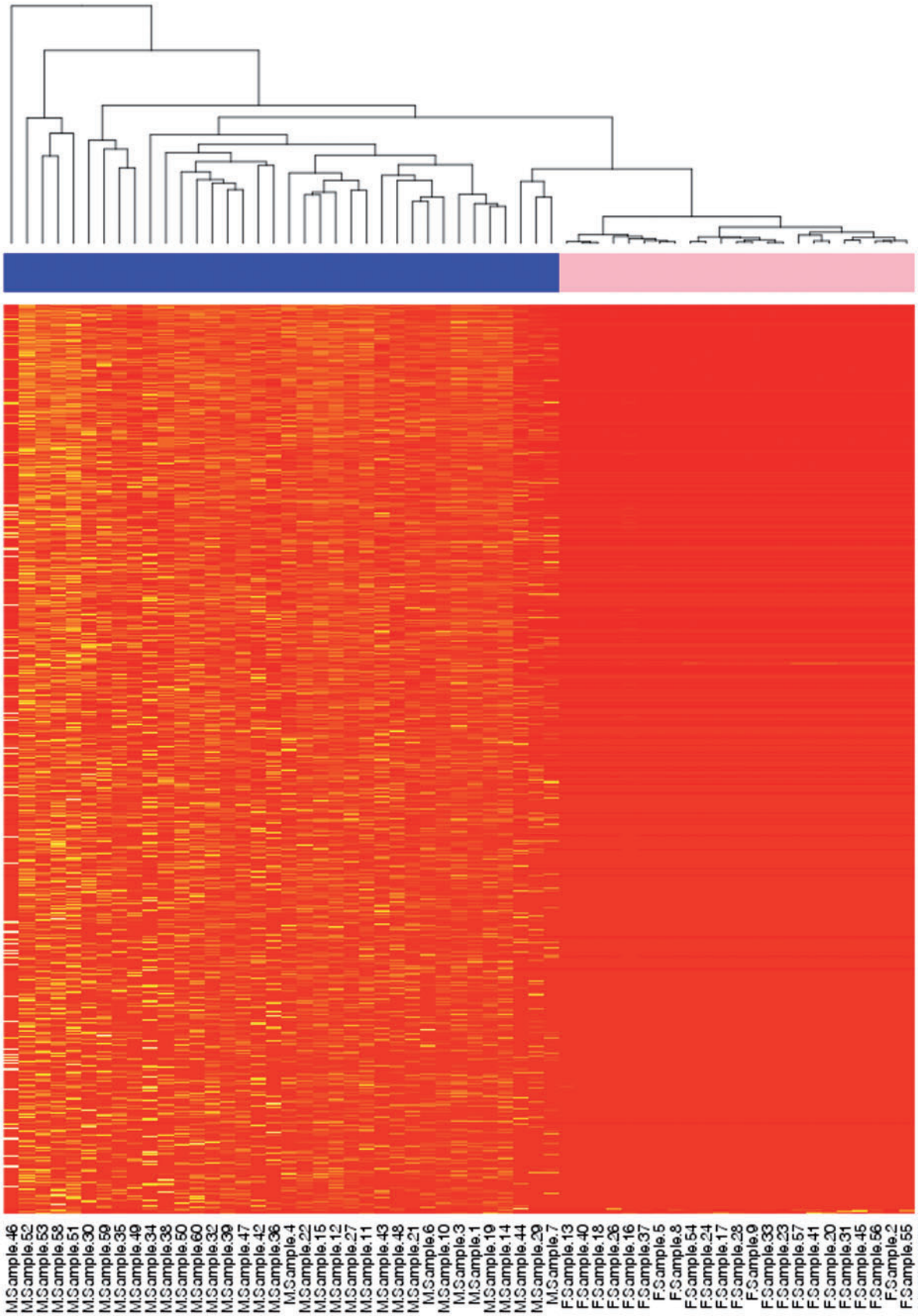

Figure 1. Unsupervised clustering of repetitive element loci expression on chromosome Y. Each column represents a sample whereas the color in each row represents expression level from a repetitive element locus on the Y chromosome. Lighter color indicates more expression. The colored boxes underneath the dendrogram denote the sex of each sample with pink representing females and blue males. In the sample name, the first letter followed by a dash also denotes sex.

Table 1. Expressed RE in the human OFC[a]

More than 20 reads in more than 3/4 of the samples					
	Sense to Exon	Sense to Intron	Anti-Sense to Exon	Anti-Sense to Intron	Intergenic
Sense in Repeat Masker	7157	3575	834	1394	5541
Anti-Sense in Repeat Masker	6898	3653	965	1029	4864
Total	14055	7228	1799	2423	10405
More than 5 reads in more than 3/4 of the samples					
Sense in Repeat Masker	9973	12187	1955	3443	9757
Anti-Sense in Repeat Masker	9982	12400	2293	2534	8790
Total	19955	24587	4248	5977	18547

[a]The columns denote the genomic compartment in which the repetitive element resides and whether the orientation of repetitive element transcription is sense or antisense to exons and introns. The rows denote the orientation of transcription for the repetitive element with respect to its annotation in the UCSC RepeatMasker hg19 annotation. Since our analysis is strand-specific, some transcripts mapped to multiple compartments and could not be assigned to a single compartment i.e. mapped to anti-sense exon or anti-sense intron and another genomic compartment. These transcripts are counted twice in this table.

Table 2. Expressed RE in human peripheral blood lymphocytes[a]

	Sense to Exon	Sense to Intron	Anti-Sense to Exon	Anti-Sense to Intron	Intergenic
Sense in Repeat Masker	4839	3088	562	807	2551
Anti-Sense in Repeat Masker	4651	3407	570	680	2331
Total	9490	6495	1132	1487	4882

[a]The columns denote the genomic compartment in which the repetitive element resides and whether the orientation of repetitive element transcription is sense or anti sense to exons and introns. The rows denote the orientation of transcription for the repetitive element with respect to its annotation in the UCSC RepeatMasker hg19 annotation. Since our analysis is strand-specific, some transcripts mapped to multiple compartments and could not be assigned to a single compartment.

Material, Table S5), more SINE elements and less LINE elements are expressed in the OFC than expected, which is consistent with our previously reported findings (10). A full listing of individual RE types and subfamilies expressed in the OFC are listed in Supplementary Material, Table S4.

As the expression threshold of more than 20 reads may be too stringent, we repeated our analysis with a cutoff of more than 5 reads per sample (Table 1). Over 70,000 RE are expressed at this lower threshold, with 34% of the expression occurring in introns. As before, the bulk of repeat loci expression occurred in the sense direction in the genes.

RE expression in peripheral blood lymphocytes and differential expression with OFC

In order to determine whether RE expression is unique to the brain, we performed RNAseq on 8 peripheral blood lymphocyte (PBL) samples collected from living individuals and processed this data using our analysis pipeline (see Supplementary Material, Table S6 for sequencing details). Based on our cutoff of at least 20 reads in three quarters of the samples, we detected expression from 22221 RE in PBL. Similar to the OFC, the majority of PBL expressed RE were transcribed in the same direction as the gene in which they were located (Table 2). In addition, the classes of expressed RE are similar to the OFC with 37% SINE elements, 24% LINE elements, 16% LTR elements, 11% simple repeats, 6% DNA transposons and 6% low complexity regions (Supplementary Material, Table S7; this table also includes information on RE subfamilies). The types of expressed RE in PBL are similar between exons, introns and intergenic regions. In addition to having more expression than expected from SINE elements and less than expected in LINE elements, PBL appears to express more LTR elements than expected (13.5%) based on the number of annotated LTR elements in the genome.

We performed differential expression analysis between PBL and OFC in order to validate our mapping approach of considering only uniquely mapping sequencing reads (mapping quality score $\geq$40). Since we only sequenced 8 PBL samples, we initially restricted our analysis to the 15 OFC samples without a psychiatric diagnosis in order to avoid problems that may arise as a result of sample imbalance between the two groups (19). Our analysis excluded RE on the sex chromosomes, RE with a base mean value of less than 10 reads or with an absolute log fold change below one. Using these criteria, we discovered 66713 differentially expressed RE between PBL and OFC with an adjusted p-value of 0.05 and an absolute log fold change of at least 1 (Supplementary Material, Table S8). We were able to confirm the results of this analysis by quantitative PCR in 5 completely new PBL RNA samples and 5 OFC RNA samples that were not included in the differential expression analysis (Fig. 2). Note that all of the differentially expressed RE loci we selected for quantitative PCR analysis validated ($P = 0.0001$ - 3.49e-06) with the exception of one locus (Fig. 2E). The RE locus at chr19:54944913 – 54945635 was close to reaching the threshold of statistical significance ($P = 0.079$) but failed to do so because of one PBL sample with a low delta-Ct value. These data suggest that RE expression levels in PBL and OFC vary between individuals. Differentially expressed RE were largely in introns (26756) and intergenic (23230) regions instead of exons (14232). However, out of the 1000 differentially expressed RE with the most significant adjusted p-value, the majority were in exons (749), while a smaller proportion were in intergenic regions (165), and few in introns (86). Although much of the differential expression between OFC and PBL appears to be plus/minus (minus defined as an average of less than 5 reads/sample), we did

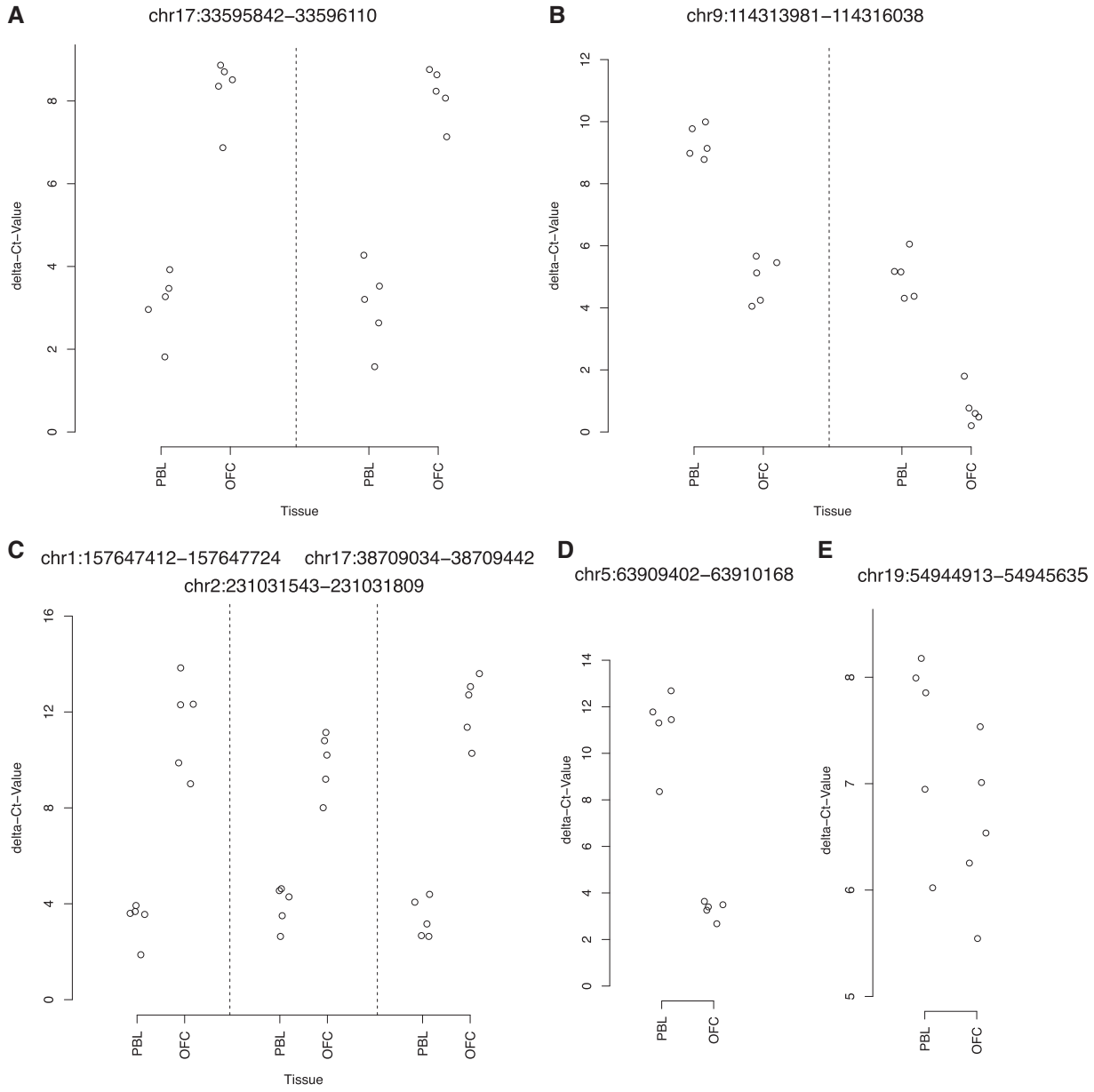

Figure 2. Quantitative PCR confirmation of differentially expressed RE between PBL and OFC. Five new PBL and five OFC samples that were not included in the differential expression analysis were used. The delta Ct-value, normalized to the GAPDH RNA levels, is plotted on the y-axis. For each RE locus in panel **A** (from left to right $P = 3.49e\text{-}06$ and $P = 1.76e\text{-}05$) and **B** ($P = 4.04e\text{-}06$ and $P = 4.64e\text{-}06$) two different PCR primer/probe pairs were designed. The similarities of the results between different quantitative PCR assays demonstrate that RE expression can be reliably measured by this method. Panels **C** ($P = 0.0001$; $P = 2.84e\text{-}05$ and $P = 4.28e\text{-}06$), **D** ($P = 0.0001$) and **E** ($P = 0.079$) show 5 additional RE differential quantitative PCR reactions.

detect loci expressed in both tissues but at differing levels i.e. we identified 8831 'downregulated' RE loci (7975 exons, 408 introns and 448 intergenic) with 50 or more reads. For all differentially expressed loci the median absolute expression change was 11.9 fold (exons: 10.2; introns: 11.1; intergenic: 17.0) and the upregulated loci had a median read count of 41 whereas the downregulated loci had a median read count of 3.3 (exons: 77.4/6.1; introns: 34.3/3.0; intergenic: 35.9/2.2). SINE elements make up the largest proportion of differentially expressed RE with 46%. LINE elements account for 21% of the differential expression, LTR elements for 12%, simple repeats for 8%, DNA transposons for 7% and low complexity regions for 5% (Supplementary Material, Table S9; this table also includes information on differentially expressed RE subfamilies). We also performed an additional analysis comparing RE expression in the 8 PBL samples to all 59 OFC samples. Due to the imbalance in sample size, we modified the requirement that RE have an average of 10 reads per sample to include all RE with 10 reads per sample in either the PBL or OFC. Increasing the sample size of the OFC samples resulted in 105742 RE reaching the threshold of significance (Supplementary Material, Table S10). However, there was extensive overlap between the two analyses. A total of 62372 (93%) of the 66713 differentially expressed RE from the original analysis, including the top 1000 most significant RE, reached significance when all OFC samples were used. As in the original analysis, differentially expressed RE were largely in introns

Table 3. RE splice junctions in humans[a]

Tissue	RE Splice Site	Acceptor Sites in Exons			Donor Sites in Exons			Unique	
		Total	Exons	Genes	Total	Exons	Genes	Exons	Genes
OFC	Intron	18961	14429	7908	19978	15062	7910	27665	10238
OFC	Intergenic	2629	1976	1510	4710	3406	2397	5264	3632
PBL	Intron	11173	9196	5700	12683	10059	5935	18355	7732
PBL	Intergenic	976	818	633	2010	1561	1175	2352	1731

[a]Splice junctions between RE and annotated exons in the orbitofrontal cortex and PBL are broken down by acceptor and donor sites. As some exons form multiple junctions with RE and contain both donor and acceptor sites, the number of unique exons and genes are listed in the last two columns.

(46197) and intergenic (38732) regions instead of exons (17108). SINE elements make up the largest proportion of differentially expressed RE with 36%. LINE elements account for 30% of the differential expression, LTR elements for 11%, simple repeats for 7%, DNA transposons for 9% and low complexity regions for 8%.

RE splice junction identification

Since our sequencing data revealed that RE are largely expressed in the sense direction of the gene in which they reside, we looked for splice junctions between annotated exons and RE located in annotated introns. As splice junctions are very short (i.e. two bases in length), the amount of sequence coverage required to detect moderate to low abundance splice junctions requires a higher depth of sequencing than that is required for differential expression analysis. Therefore, we decided to include all OFC samples in our analysis to identify RE splice junctions. By analysing the junction.bed files generated by the Tophat2 (20) short read aligner we were able to identify the presence of 38939 such splice junctions in the orbitofrontal cortex (Table 3, Supplementary Material, Table S11 – the psychiatric diagnosis of each sample is included in the supplementary table). The intronic RE was the splice donor in 18961 and the splice acceptor in 19978 instances. These donor and acceptor sites map to 16488 and 17030 individual RE loci respectively. They form splice junctions with splice acceptor sites in 14429 different annotated exons in 7908 genes and splice donor sites in 15062 different annotated exons in 7910 genes. When we only consider unique exons and genes we find that RE splicing in the OFC occurs with 27665 exons in 10238 genes. The existence of the identified splice junctions was verified using PCR analysis (Fig. 3) in a subset of junctions. Note that we have had greater than 90% success in validating RE splice junctions via PCR. We then investigated the consensus sequences around the splice sites and discovered that RE-exon splice junctions overwhelmingly occurred at canonical GT-AG sites (Fig. 4, Supplementary Material, Table S12).

We repeated our analysis in intergenic regions and discovered 7339 intergenic RE that formed splice junctions with annotated exons (Table 3, Supplementary Material, Table S11). In 4710 instances, the intergenic RE were the splice acceptors and in 2629 instances they were the splice donors. The acceptor sites were located in 3532 RE and the donor sites were located in 2078 RE. These RE formed splice junctions with splice donor sites located in 3406 annotated exons in 2397 genes and with acceptor sites in 1976 annotated exons in 1510 genes. Overall, intergenic RE splicing encompasses 5264 unique exons in 3632 different genes.

Since the 59 OFC samples included individuals with schizophrenia, bipolar disorder and major depression, we performed a differential expression analysis using the number of reads in each sample spanning the junctions to determine whether any of the donor or acceptor sites are associated with a psychiatric disorder. None of the 46278 donor and acceptor sites from Supplementary Material, Table S11 were differentially expressed in any of the three disorders; while the unadjusted p values ranged from 0.00004 to 1 all p values were 1 after multiple testing correction (Supplementary Material, Table S13).

RE splicing is not restricted to the brain and we identified 23856 splice junctions between annotated exons and intronic RE in our PBL sequencing data (Table 3, Supplementary Material, Table S14). To confirm the existence of RE splicing in PBL and in intergenic regions, we performed additional PCR reactions and Sanger sequencing (Fig. 5). The RE was the splice donor in 11173 instances and the splice acceptor in 12683 instances. These RE formed splice junctions with splice acceptor sites in 9196 annotated exons in 5700 genes and splice donor sites in 10059 annotated exons in 5935 genes. When we only consider unique exons and genes, we find that RE splicing in PBL occurs in 18355 exons and 7732 genes. We also identified 2986 (976 donor/2010 acceptor) intergenic RE splice junctions in PBL. These junctions consisted of acceptor sites in 818 exons (633 genes) and donor site in 1561 exons (1175 genes). In total there were 1731 unique genes and 2352 unique exons that form splice junctions with intergenic RE in PBL.

We wanted to determine the correlation between RE splicing and RE expression in the orbitofrontal cortex. We find that out of the 8032 intronic RE expressed in the orbitofrontal cortex based on a cutoff of 20 reads in ¾ of the samples and a strand free annotation, 1147 (14.3%) contained splice junctions. When we repeat our analysis using a threshold of 5 reads in ¾ of the samples and a strand free annotation, we classify 27609 intronic RE as expressed and out of these 4376 RE contain a splice junction (15.8%). Similarly, at the 20 read threshold out of the 7336 intergenic RE that were classified as expressed 246 contained splice junctions (3.4%). At the 5 read threshold, 980 out of 18547 intergenic OFC expressed RE contained splice junctions (5.3%).

The location of RE splicing within the transcript

In order to determine whether RE splicing preferentially occurred in UTRs or coding regions, we determined the location of donor and acceptor sites in gene exons. In the OFC, intronic RE donor and acceptor sites primarily form splice junctions with exons located in coding regions (Table 4) accounting for approximately 73% of all splice junctions. A large number of splicing events were also observed in 5 prime UTR exons (~24%) but relatively few RE splicing

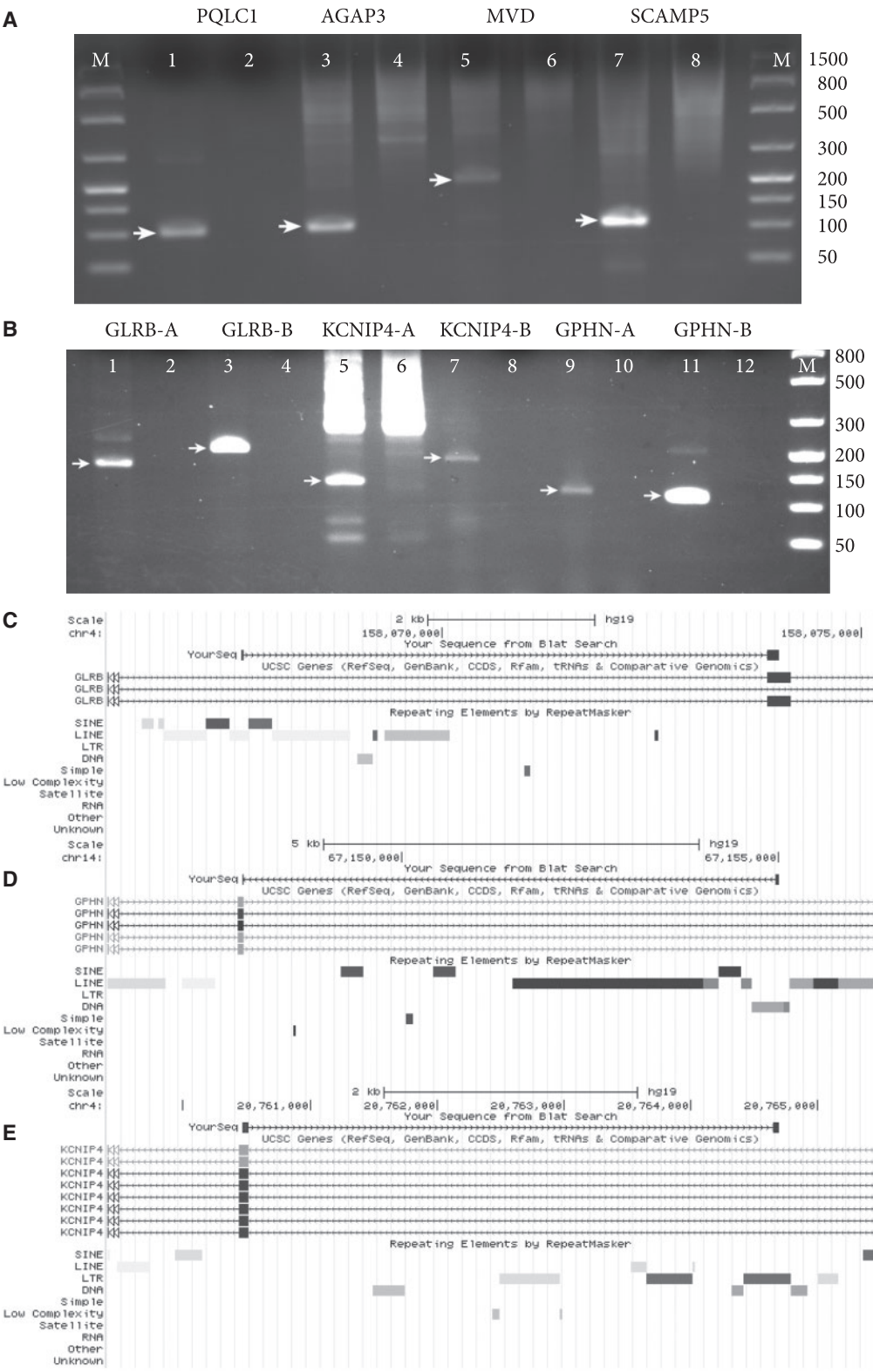

Figure 3. PCR validation of intronic RE splice sites in OFC. In **(A)** and **(B)** the same primer pair was used to amplify brain cDNA (odd numbered lanes) and genomic DNA (even numbered lanes). The arrows show the PCR bands that were isolated and sequenced for confirmation. The name of the gene RE splices into is noted on top. In **B**) we selected RE that have at least two splicing events and may potentially introduce an exon and an open reading frame. **(C, D, E)** UCSC alignment of splice junctions between RE and exons of three different genes corresponding to lanes *GLRB*-A, *KCNIP4*-A, and *GPHN*-A in **B**. The top track labelled 'YourSeq' is the alignment of the sequenced PCR amplicon. The rectangular boxes in this track represent the sequence of the amplicon and lines are used to represent gaps in the alignment. The tracks below 'YourSeq' show annotated genes and transcripts in the human genome (build hg19). In these tracks exons are represented as rectangular boxes and introns are represented as lines. The bottom of each figure contains the repeat masker track. Note that one side of 'YourSeq' overlaps a RE whereas the other end overlaps an annotated exon.

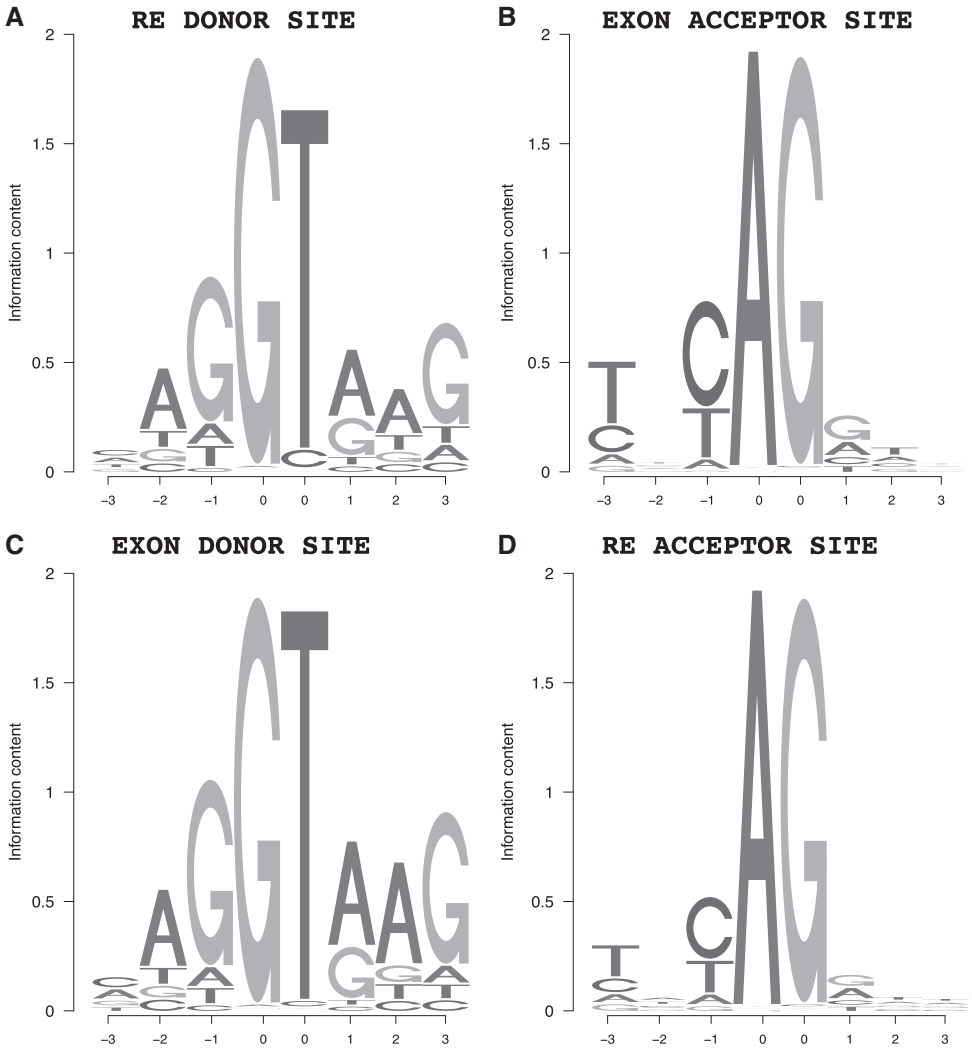

Figure 4. Sequence of intronic RE splice junctions in the orbitofrontal cortex. The x axis denotes the position of the nucleotide from the splice junction, which is labelled as zero. The height of the nucleotide represents the relative frequency of a base at a particular position. Panel **A)** shows the sequences at splice junctions between RE containing donor sites and the corresponding exon acceptor sites Panel **B)**. Panel **C)** shows the sequences at splice junctions between exons containing donor sites and corresponding RE containing acceptor sites Panel **D.**

events were detected in the 3 prime UTR (∼3%). A very similar distribution of RE splicing is observed in human PBL in which splice junctions are formed in acceptor and donor sites located in coding exons (∼74%) followed by 5 prime UTR (∼23%) and the 3 prime UTR (∼3%). The number of acceptor and donor splice sites located in coding exons were approximately equal in both the OFC (36% and 37% of all sites) and PBL (35% and 39% of all sites). However, in 5 prime UTR exons acceptor sites occur approximately 1.4 times more frequently than donor sites in the OFC and 1.2 times more often in PBL. In 3 prime UTR exons donor sites occur approximately 2.8 times more frequently than acceptor sites in the OFC and 3 times more often in PBL. Intergenic RE form junctions with acceptor or donor sites in exons that are most frequently located in coding regions (59% in OFC; 58% in PBL), followed by the 5 prime UTR (23% in OFC; 20% in PBL) and the 3 prime UTR (18% in OFC; 21% in PBL). Exons in coding regions that form splice junctions with intergenic RE contain more donor sites (2.1x in OFC; 2.7x in PBL). Similarly, acceptor sites in 5 prime UTR exons were 4 times more frequent in both OFC and PBL whereas exons in the 3' UTR almost exclusively contain donor sites. However, we find that donor sites in 5 prime UTR exon sometimes form splice junctions with acceptor sites in intergenic RE that are downstream of the gene and in certain instances, donor sites in intergenic RE upstream of a gene can form a splice junction with an acceptor site in an exon located in the 3 prime UTR.

When we repeated our analysis ignoring the strand information for annotated exons, we found twice as many 5 prime UTR exons containing donor sites (additional 329 in OFC and 105 in PBL) that form splice junctions with acceptor sites in intergenic RE (Supplementary Material, Table S15). Similarly, we found 7 times and 14 times more acceptor sites in 3 prime UTR exons in the orbitofrontal cortex (additional 210) and PBL (additional 203) respectively that form junctions with donor sites in intergenic RE. These junctions likely belong to unannotated antisense transcripts. Our analysis also identified a modest number of potential antisense transcripts in every other compartment (Supplementary Material, Table S15).

RE types involved in splicing

To assess whether certain RE classes are more prone to splicing in humans, we calculated the frequency at which each RE class

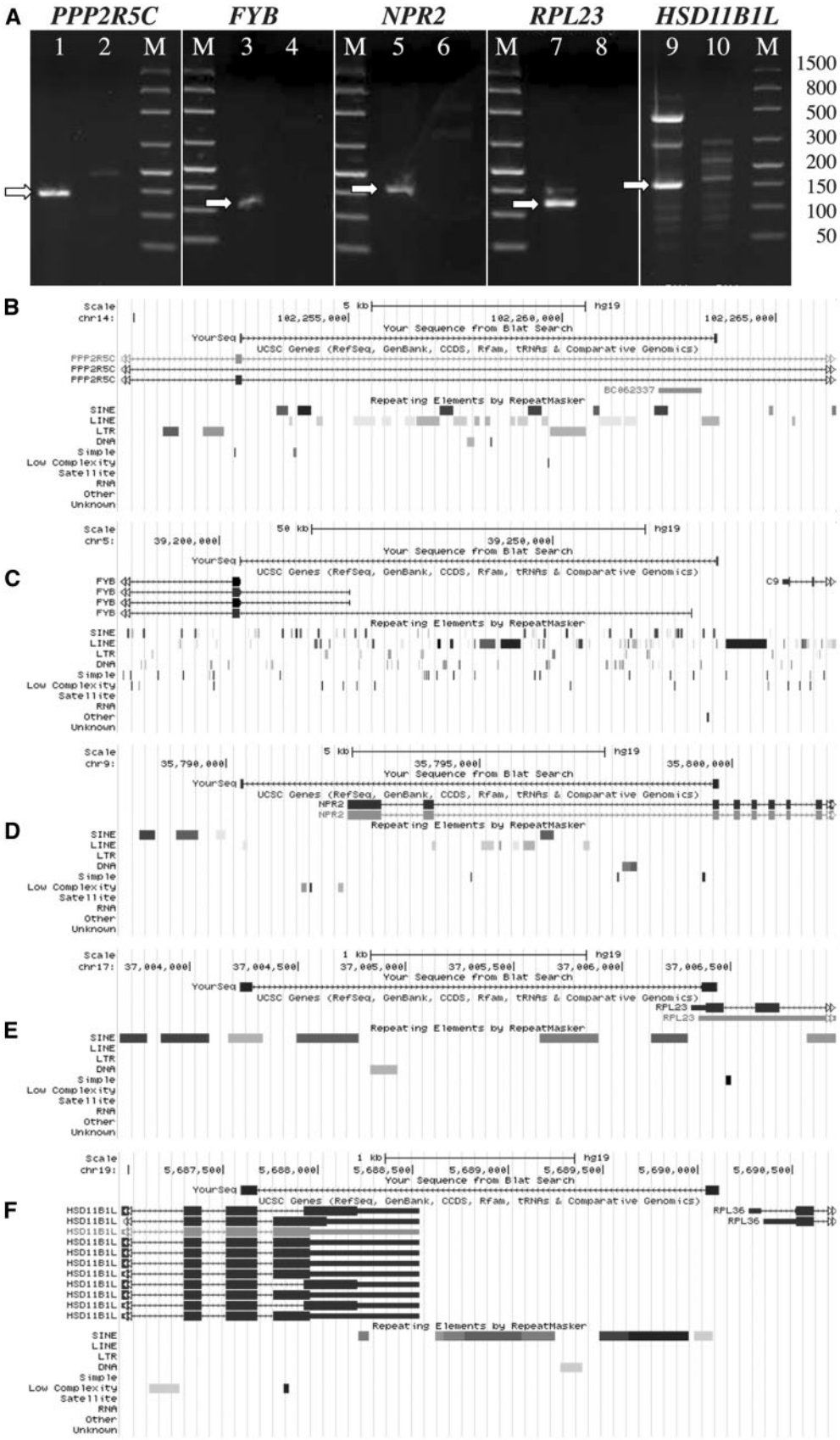

Figure 5. PCR validation of RE splicing in PBL and intergenic regions. In (**A**), the same primer pair was used to amplify cDNA (odd numbered lanes) and genomic DNA (even numbered lanes). Lanes 1, 3, 5 and 7 show amplification of cDNA made from commercially available PBL RNA. Lane 9 shows amplification of cDNA from

Table 4. Location of RE splice junctions in human mRNAs[a]

Tissue	RE Location	5' UTR		CDS		3' UTR	
		Type of Site in Exon		Type of Site in Exon		Type of Site in Exon	
		Acceptor	Donor	Acceptor	Donor	Acceptor	Donor
OFC	Intron	5499	3820	14236	14718	311	864
PBL	Intron	3144	2691	8860	9762	171	518
OFC	Intergenic	992	254	1043	2203	33	966
PBL	Intergenic	398	95	380	1034	8	506

[a]The mRNA location of annotated exons that form splice junction with RE in the human orbitofrontal cortex and PBL are shown. As certain splice junctions mapped to multiple transcripts, they may have mapped to both a UTR and a coding region and may have been counted twice.

forms splice junctions (Supplementary Material, Table S16). In both the OFC and PBL, LINE elements accounted for the largest number of splice junctions (OFC 36.7%/PBL 38.1%) for intronic RE but SINE elements were a close second (OFC 35%/PBL 33.1%). In contrast, LTR elements were found to form the most splice junctions in intergenic RE (OFC 36%/37% PBL). When we took into account the length and the number of annotated RE loci in the genome for each RE class, we discovered that LTR and DNA elements form splice junctions more frequently whereas LINE elements form splice junctions less frequently than expected (Fig. 6).

We wanted to determine whether the RE splice junctions we identified originated from younger RE that are potentially capable of transposition. Since we used split reads in which one portion of the read aligns to an exon, a non-repetitive region, it is technically possible that the splice junctions we identified might be located in young RE. Thus, we overlapped the location of the RE splice variants we identified with the location of known RE polymorphisms in the hg19 build of the human genome (21). Out of the 1593 RE that have insertional polymorphisms, our splice junctions only overlapped 3 SVA elements (2 intronic and 1 intergenic, dbRIPID: 3000062, 3000060, 3000031) suggesting that the vast majority of the splicing we are detecting occurs in older RE.

Association of RE expression with transcriptional start sites

We then wanted to determine the overlap between RE splicing and previously published Cap Analysis Gene Expression (CAGE) data (5) that identified 117,165 RE loci that provide alternative transcriptional start sites to human RNAs. These loci are compiled from CAGE sequencing performed in the human brain, urogenital, adipose, fibroblast, liver, digestive tract, hepatoma, neuroblastoma, monocytic leukaemia, testis, monocyte and embryonic tissues. Out of the 7376 intergenic splice sites we identified, 33 RE acceptor and 39 RE donor sites overlapped with the RE reported in the CAGE data. Similarly, out of the 38939 intronic repeat splice junctions in the orbitofrontal cortex 238 splice acceptors and 389 splice donors overlapped with the RE loci in CAGE data. In addition, we found 2277 out of the 31351 OFC expressed RE (Table 1) to overlap with the CAGE RE loci.

RE splicing in the brain across species

We then examined RE splicing in the mouse, rat, zebrafish, drosophila, dog and rabbit brain tissue using publicly available RNA-seq data (see Methods and Supplementary Material, Table S17). This data varied in terms of sequencing depth, read length and whether it was single or paired end sequencing (Supplementary Material, Table S17). Compared to the human sequencing data we generated, the sequencing depth in the non-human samples was much lower. The average number of aligned read in OFC was approximately 134 million whereas in the non-human sequencing the number of aligned reads ranged between 11 to 45 million (Supplementary Material, Table S17). Unlike the OFC study, the number of samples in the non-human study ranged from 8 to 15. In addition, the non-human sequencing data was not directional. However, the non-human sequencing was also performed on the Illumina platform. Donor and acceptor sites between intronic/intergenic RE and annotated exons were observed in every species examined (Table 5, Supplementary Materials, Table S18–S23). In all species, intronic RE formed splice junctions with exons located in coding regions (59%-91% of RE-exon junctions), followed by exons in the 5 prime UTR (8%-26% of RE-exon junctions) and the 3 prime UTR (1-15% of RE-exon junctions) (Table 6). Intergenic RE formed more splice junctions with exons located in UTRs (13%-33% 5' UTR; 9%-38% 3'UTR) but the majority of the splicing occurred in coding regions (49%-61%). Fewer RE splice junctions were detected in the dog and rabbit brains likely because the number of annotated transcripts in these species is much lower compared to rats, mice or humans. Since the sequencing methods for each study varied greatly, we are unable to compare the absolute number of splicing events in an RE class between species. However, we can estimate the frequency at which a RE class forms splice junctions with annotated exons within a species (Supplementary Material, Table S24). In mice, we find LINE (31%) and SINE (30%) followed closely by LTR (27%) elements in introns to form the most splice junctions whereas in rats, intronic SINE elements (51%) account for the vast majority of RE splicing (Supplementary Material, Table S24). In contrast, intronic DNA transposons (69%) are the prominent RE class involved in forming splice junctions in zebrafish. Since the relative abundance of each type of RE in the genome can vary across species (we did not adjust for this in our analysis), the

commercially available total brain RNA. The arrows show the PCR bands that were isolated and sequenced for confirmation. The name of the gene each RE splices into is noted on top. (**B**) UCSC alignment of splice junction between an intronic RE expressed in PBL and an exon of *PPP2R5C* in PBL. (**C,D,E**) UCSC alignment of splice junctions between intergenic RE expressed in PBL and exons of *FYB*, *NPR2* and *RPL23*. (**F**) UCSC alignment of splice junctions between an intergenic RE expressed in OFC and an exon of *HSD11B1L*. The top track labelled 'YourSeq' is the alignment of the sequenced PCR amplicon. The rectangular boxes in this track represent the sequence of the amplicon and lines are used to represent gaps in the alignment. The tracks below 'YourSeq' show annotated genes and transcripts in the human genome (build hg19). In these tracks exons are represented as rectangular boxes and introns are represented as lines. The bottom of each figure contains the repeat masker track. Note that one side of 'YourSeq' overlaps a RE whereas the other end overlaps an annotated exon.

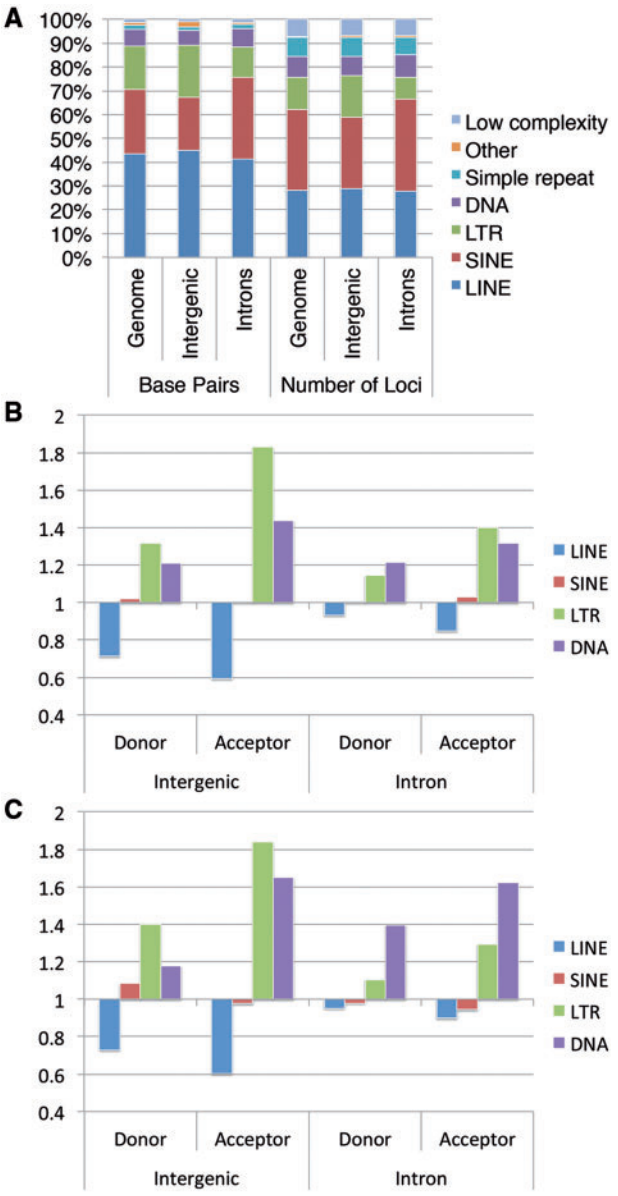

Figure 6. Abundance of splice junctions in each class of RE relative to RE annotation. Panel (**A**) shows the relative abundance of each class of RE annotated in various genomic compartments. The actual number of splice junction donor and acceptor sites in the orbitofrontal cortex (**B**) or PBL (**C**) that are in each class of RE is divided by the number that would be expected if all bases within REs were equally likely to form splice junctions. The four most abundant classes of REs (which make up more than 95% of total RE base pairs) are shown.

frequency with which each type of RE forms splice junctions may be driven at least in part by the relative abundance of RE types in each genome.

Expression and coding potential of RE splice variants

We discovered 4802 potential exonization events in which an RE contained both an acceptor and a donor site. We only considered potential exons that added at least 20 bases to the gene transcript. We counted an exonization event as being unique if any of the donor and acceptor sites differed in the RE or the flanking exons. Of the 4802 potential exonization events, RE potentially form 3900 distinct (non-overlapping) exons in 2289 different genes. PCR analyses performed in Fig. 3B validate the presence of acceptor and donor splice sites in three such RE that are potentially exonized.

In order to determine the abundance of these splice variants, we calculated read coverage at the putative exonized RE and the flanking exons. Out of a possible 4802 putative exonization events, in 31 instances, the average coverage for the putative exon was higher than both its surrounding exons, in 234 instances it was higher than only one of the exons and in 4537 instances it was expressed at lower levels than the two surrounding exons (Supplementary Material, Table S25). The putative exons on average had 9.5% (2.9% median) of the read coverage of the surrounding exons. Hence, most of these putative exons give rise to minor splice variants. In order to validate this finding, we performed quantitative PCR comparing relative abundance between the canonical *GLRB* transcript and an exon/RE splice junction in the same gene (Fig. 7). We performed a standard curve analysis to ensure that each quantitative PCR assay efficiently amplified its target. The PCR assay targeting the canonical *GLRB* transcript had an amplification efficiency (slope) of -3.49 whereas the PCR assay targeting the RE splice junction had an amplification efficiency of -3.54 as measured by the SDS 2.4 software from Applied Biosystems. The RE variant in the 5 OFC samples tested on average had a Ct value that was 6 cycles higher and this difference was statistically significant ($P = 2.85e\text{-}05$), which is consistent with our sequencing analysis.

We characterized the impact of RE splicing into coding regions by examining the coding potential of the 4802 novel transcripts that may arise. We discovered that 4263 out of the 4802 putative exonization events we identified formed a splice junction with at least one exon located in a coding region. Because putative exonization events with different donor and acceptor sites in the annotated exons were counted separately, they represent the insertion of 3906 different putative RE-derived exons. The RE-derived exons range in length from 20 to 4926 nucleotides in length with a median length of 98 nucleotides, a mean length of 115 nucleotides and a standard deviation from the mean of 127 nucleotides. We translated these putative exons in all three frames (frame 1,2 and 3 for exons on the plus strand and frame -1,-2 and -3 for exons on the minus strand). We found that 2487 out of the 3906 putative exons (64%) had an open reading frame in at least one of the three translated frames (Supplementary Material, Table S26).

We then characterized the transcript structure resulting from the putative RE exonization event for the *AGAP3* mRNA by performing 5 prime and 3 prime rapid amplification of cDNA ends (RACE) reactions (Fig. 8 and Supplementary Material, Data S1). Primers targeting the RE exon were designed and 5 and 3 prime regions of the transcripts were PCR amplified, cloned and subjected to Sanger sequencing. The resulting RACE products, especially the 5 prime RACE product which starts in a highly GC rich region, may not represent the full-length transcript. In addition, we cannot discount the possibility that the five prime and 3 prime amplicons might be from different transcripts. The obtained RACE amplicons show that the RE is spliced into the canonical long gene transcript and the expected exon structure prior to the RE splicing event is retained (Fig. 8). The 3 prime RACE product starts in an RE but the splice junction with the canonical exon is located downstream in the intron. The obtained transcripts have 100% homology with the reference human genome (hg19) at the nucleotide level. In addition, RE splicing appears to occur in frame assuming that translation initiates at the canonical translation start site. *In silico* translation of the 5 prime RACE amplicon reveals that RE-derived nucleotides

Table 5. RE splice junctions in the brain of non-primate species[a]

Tissue	RE Splice Site	Acceptor Sites in Exons			Donor Sites in Exons			Unique	
		Total	Exons	Genes	Total	Exons	Genes	Exons	Genes
Mouse	Intron	6177	5306	3865	6040	5189	3708	10165	5590
Mouse	Intergenic	502	430	357	779	664	512	1080	849
Rat	Intron	1543	1392	1234	1572	1429	1240	2785	2051
Rat	Intergenic	332	176	166	272	161	145	328	303
Zebrafish	Intron	664	604	533	497	476	421	1069	850
Zebrafish	Intergenic	146	88	146	121	102	94	219	200
Drosophila	Intron	238	195	163	154	125	105	314	234
Drosophila	Intergenic	67	51	49	75	58	57	109	104
Dog	Intron	157	140	123	125	110	99	248	183
Dog	Intergenic	30	24	23	27	22	21	46	43
Rabbit	Intron	56	51	46	58	54	49	104	84
Rabbit	Intergenic	11	11	10	10	10	10	21	20

[a]The number of RE donor and acceptor sites identified in RNAseq data generated in the brain of various non-primate species is summarized. The total number of exons and genes involved are listed.

Table 6. Location of RE splice junctions in non-primate mRNAs[a]

Tissue	RE Location	5' UTR		CDS		3' UTR	
		Type of Site in Exon		Type of Site in Exon		Type of Site in Exon	
		Acceptor	Donor	Acceptor	Donor	Acceptor	Donor
Mouse	Intron	1721	1232	4630	4540	67	200
Rat	Intron	244	200	1266	1273	11	23
Zebrafish	Intron	124	64	469	391	29	18
Drosophila	Intron	54	29	110	79	32	19
Dog	Intron	17	6	142	118	1	2
Rabbit	Intron	8	2	48	53	1	1
Mouse	Intergenic	231	21	176	301	7	213
Rat	Intergenic	103	4	77	87	0	50
Zebrafish	Intergenic	38	2	47	34	9	52
Drosophila	Intergenic	10	1	16	26	10	22
Dog	Intergenic	12	1	16	11	0	4
Rabbit	Intergenic	5	0	4	5	1	2

[a]The number of RE donor and acceptor sites identified in RNAseq data generated in the brain of various non-primate species is summarized. The total number of exons and genes involved are listed.

located at the end of the sequence introduce the amino acids LMNRR onto the canonical *AGAP3* protein (NP_001036000.1) (Supplementary Material, Data1). Similarly, the 3 prime RACE product encodes an open reading frame that begins with the amino acids RGLAGPARVRAQASPSPSGSFSVTPLPRGKF which are in a frame with the canonical end of the *AGAP3* protein (NP_001036000.1). The amino acids 32-61 encoded by this amplicon are homologous to predicted *AGAP3* proteins XP_016867224.1 and XP_005250000 (Supplementary Material, Data1).

Association of RE splice variants with translating ribosomes

In order to determine whether RE splice variants are translated into proteins, we analysed previously published Translating Ribosome Affinity Purification sequencing (TRAP-seq) data (22) from the mouse cerebellum. This data set consisted of TRAP-seq (51bp paired end reads) performed on Purkinje cells (pj), granule cells (gc) and Bergmann glia (bg). Similar to the RNA-seq data, the TRAP-seq results contained numerous acceptor and donor sites in intronic (the acceptor and donor sites in pj:3226/2612; bg:826/691; gc:2685/2260) and intergenic (pj: 629/335; bg:161/92; gc:453/201) RE that formed splice junctions with annotated exons in thousands of different genes (pj:3113; bg:1115; gc:2768) (Supplementary Material, Table S27). Since bg samples were sequenced at lower depth (mean read counts/sample pj:97,764,335.75; bg:31,762,644; gc:98,668,986.75), the observed differences in the number of splice junctions between cell types is likely an artefact. These mRNAs largely had splice junctions that were between intronic RE and annotated exons located in coding regions (Table 7). In addition, we were able to detect individual RE that contained both donor and acceptor sites in every cell type (pj:195; bg:34; gc:138). When we considered RE 'exons' that form a splice junction with at least one canonical exon located in a coding region, we found that their potential to extend the encoded protein was similar to what was observed in the human orbitofrontal RNA sequencing data (pj:80/137(58%); bg:15/21(71%); gc:50/92(54%)) (Supplementary Material, Table S28).

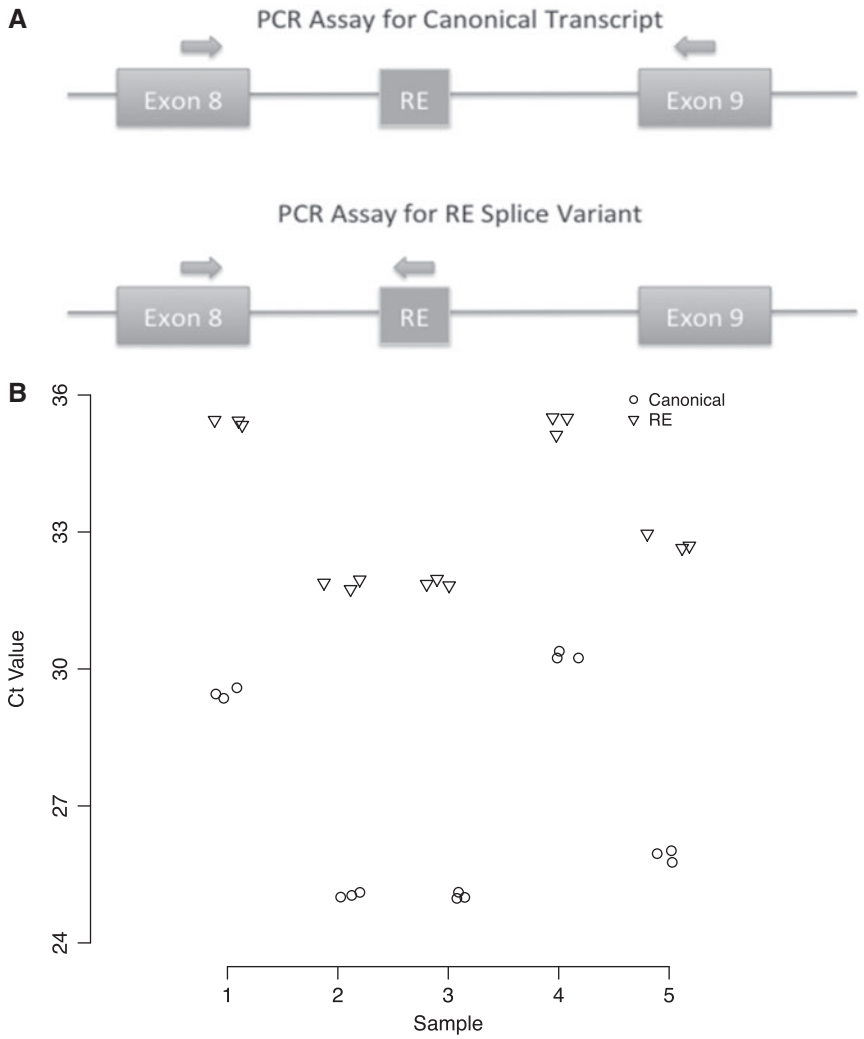

Figure 7. Quantitative PCR comparing levels of the canonical and RE spliced GLRB transcripts. Panel (**A**) summarizes the PCR strategy. The arrows represents the location of the PCR primers for each assay. Panel (**B**) plots the Ct values of the quantitative PCR reactions, which were performed in triplicate. The difference in expression between the canonical and the RE splice variant is statistically significant ($P = 2.85e\text{-}05$).

Discussion

In this work, we demonstrate that RE expression is more complex than previously envisioned. We find that many older RE elements, which are fixed in the genome, are transcribed in the same direction as the gene in which they are located. These RE are likely incorporated into pre-mRNAs and the corresponding gene drives their expression. In addition, we find thousands of donor and acceptor sites in intronic and intergenic RE that form splice junctions with annotated exons. RE form splice junctions most frequently with exons located in coding regions, followed by exons in the 5 prime and then the 3 prime UTR. We find that RE splicing is not restricted to a single cell type, tissue or species and it is possible that RE splicing may be present in all eukaryotic cells. RE splicing, which affects thousands of genes, usually gives rise to minor transcript variants. Our data also suggest that there are many previously undetected RE exonization events in the human OFC. These putative exons have the potential to alter the encoded protein and may be translated.

We demonstrate that RE expression can be estimated accurately in the human genome. Our analysis of Y-chromosome RE expression in females found that very few RNA-seq reads were erroneously mapped. Based on this analysis, we defined OFC expressed RE as those loci that had more than 20 reads in more than ¾ of the samples. This stringent threshold, which the majority of the loci examined in the cluster analysis for the Y-chromosome (Fig. 1) fail to meet, gives us confidence in the accuracy of the RE we classified as expressed. Further confirmation of our mapping strategy comes from our ability to identify differentially expressed RE between OFC and PBL that were validated using quantitative PCR.

Regardless of the thresholds we employed, the majority of RE expression was found to be in genic regions. Since RE expression largely occurs in the same orientation as the genes in which they are located, the corresponding gene may regulate the expression of most RE. Given that we find nearly an equal number of RE loci that are expressed in the sense and the antisense orientation (based on the RE annotation in repeat masker) (Table 1 and Table 2), the fact that RE are largely expressed in the sense direction to genes cannot be explained by the underlying DNA structure. These findings compelled us to examine whether RE are spliced into gene transcripts. We found that splice junctions between RE and annotated exons are prevalent

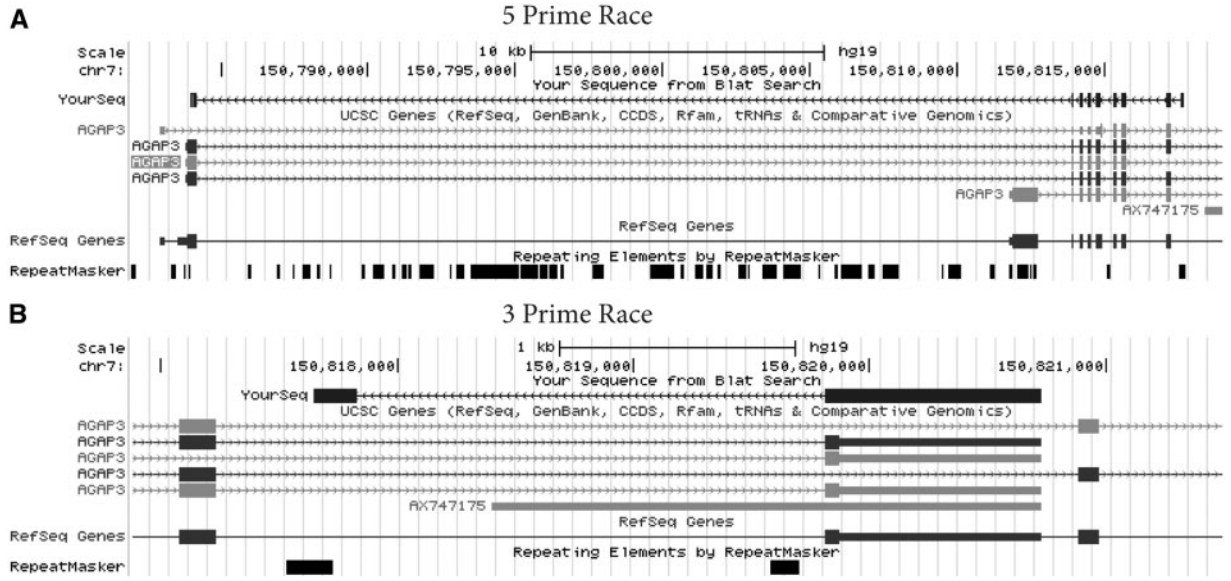

Figure 8. The structure of *AGAP3* transcript containing an RE splice junction. Panel (**A**) shows the UCSC genome track alignment of the 5 prime RACE product whereas panel (**B**) shows the same for the 3 prime RACE product. The 'YourSeq' track represents the RACE product. Both RACE products show the expected exon order with the exception of the RE splicing event. Note that the 3 prime RACE product extends into the intron prior to forming a splice junction with the annotated exon.

in coding regions of every sample we examined. When we overlapped the RE splice junctions and the expressed RE identified in the human OFC, we discovered that 14.3% of intronic and 3.4% of intergenic RE contained a splice junction. Lowering the threshold for RE expression enabled us to map additional splice junctions but did not significantly change the proportion of expressed RE associated with splicing (intronic RE 15.8%; intergenic RE 5.4%). Although we selected polyadenylated RNAs for sequencing, which are primarily mature transcripts, these results suggest that many of the expressed RE we identified may be in unprocessed pre-mRNAs. However, some expressed RE without splice junctions may also represent alternative transcription start sites, termination sites, or intron retention events in mature mRNAs and we cannot rule out the possibility that some expressed RE are independently transcribed. It is also possible that the inability of *Taq* DNA polymerase to amplify certain repetitive sequences, limitations of alignment software to accurately map RE splice sites and the fact that our sequencing data is single ended prevented us from finding additional RE splice junctions.

RE splicing gives rise to a substantial number of transcript variants in the human OFC, PBL and potentially other tissues. RE splicing was also detected in the brain of every non-human species we examined including mouse, rat, zebrafish, drosophila, dog and rabbit. The number of such splice junctions in the brain differed greatly between species. This was likely because the annotation and completeness of the genome build varied between the species i.e. the dog and the rabbit genome only have around 1500 annotated transcripts, compared to >50000 in humans. Additional factors such as differences in sequencing depth, read length and whether the sequencing was paired end also contributed to these differences. Regardless, RE splicing was more prominent in coding regions in all species and in human PBL. The fact that RE splicing is present in multiple species suggests that this biological process might be beneficial to the organism i.e. perhaps by increasing the diversity of splice variants. Based on these findings, we conclude the RE splicing is a normally occurring event. Since our data contains samples from psychiatric subjects, we performed differential expression analysis to determine whether some of the RE junctions we identified may be associated with disease. Consistent with previous findings that only subtle changes to expression are present in psychiatric disorders, none of the RE splice junctions were differentially expressed in disease. Making such an association would require a larger cohort, replication studies and higher depth of sequencing. Nevertheless, we report the diagnosis of each sample in our RE splice junction count tables (Supplementary Material, Table S11).

Our work differs from previous findings in which RE were reported to provide a large number of alternative transcription start sites for human genes (5). It is not surprising that we did not find extensive overlap between the CAGE data and our OFC RE expression or RE splicing data, given that our RNA sequencing approach was not restricted to transcriptional start sites. In addition, RNA sequencing lacks the sensitivity of CAGE in identifying transcriptional start sites. Our findings also demonstrate that Alu exonization (7) is only a subset of RE splicing that occurs in the cell. We find that LINE elements form splice junctions slightly more frequently than SINE elements in humans and mice. Our data suggest that the frequency at which an RE class is spliced relative to the other classes of RE is species specific with mice having significantly more splice junctions in LTR and zebrafish having more splice junctions in DNA elements than other RE classes. Potentially, the RE splicing we describe may be associated with the process of exonization (7) and may explain the high number of exons in humans that evolved from RE sequences.

Gage and colleagues have hypothesized that transposition is an integral part of the developing nervous system that generates cellular heterogeneity within the brain (23). More recently, Bundo et al. have reported that L1 insertions are increased in schizophrenia brain samples (24). Since transposition generates identical copies of an element, our approach of estimating expression levels based on uniquely mapping reads is not suitable for studying transposition-capable elements. In addition, our splicing analysis does not detect junctions between younger

Table 7. RE splice variants in translating ribosomes[a]

Tissue	RE Location	5' UTR		CDS		3' UTR	
		Type of Site in Exon		Type of Site in Exon		Type of Site in Exon	
		Acceptor	Donor	Acceptor	Donor	Acceptor	Donor
Purkinje	Intron	706	521	1814	2464	63	152
Bergmann Glia	Intron	209	167	441	584	25	29
Granule Cells	Intron	610	524	1628	2009	57	116
Purkinje	Intergenic	93	29	116	274	11	150
Bergmann Glia	Intergenic	30	10	30	66	4	33
Granule Cells	Intergenic	41	20	56	187	21	117

[a]RE Splice Variants associated with Translating Ribosomes are listed based on where the splice junction is located within the mRNA. As certain splice junctions mapped to multiple transcripts, they may have mapped to both a UTR and a coding region and may have been counted twice.

LINE elements and annotated exons. Potentially, our findings represent a biological process that is largely associated with older LINE elements and unaffected by actively transposing LINE elements. Alternatively, our analysis pipeline may be unable to efficiently detect splice junctions formed between canonical exons and active LINE elements.

Our findings raise the possibility that RE may generate novel functional protein isoforms. Our *in silico* analysis suggests that putative exons generated by RE splicing have the potential to alter the sequence of the encoded protein. Further evidence for this possibility was provided by RACE reactions in which we discovered that RE splicing in the AGAP3 mRNA results in a transcript that contains additional in frame amino acid sequences. Consistent with these findings is the fact that many RE splice variants, in which a donor or an acceptor site in the RE forms a splice junction with an exon located in the coding region, are associated with translating ribosomes in the mouse brain. Further characterization of the proteins encoded by RE splice variants in different species is required to establish whether RE splicing plays a functional role in the cell. We note that a significant number (37%) of exonized RE will introduce a premature termination codon into the transcript since they do not contain any open reading frames. In addition, many of the exonized RE that have the potential to extend the encoded protein may not splice in the appropriate reading frame, resulting in the introduction of premature termination codons. The presence of thousands of transcripts containing RE that encode truncated proteins alone is significant as it suggests that the number of defective transcripts present in the cell has been underestimated. These transcripts will likely trigger nonsense mediated decay and similar mechanisms, which can protect the cell from truncated proteins that are deleterious (25,26). However, given that the cell expends substantial resources to generate and translate large numbers of RE splice variants, it is unlikely that this process would have persisted during evolution if it were not somehow beneficial. Our current hypothesis is that despite giving rise to many defective transcripts, RE splicing will generate some novel functional protein isoforms.

In summary, we provide a thorough characterization of RE expression in the OFC and detail its complexities. RE splicing gives rise to a substantial number of transcript variants in OFC, PBL and potentially other tissues in a variety of organisms. Splicing of RE likely impacts gene activity, as it occurs predominantly in coding regions of mRNAs. Potentially, RE splicing may lead to pathology in conditions where defects in the splicing or RNA quality control machinery exist. On the other hand, RE splicing may generate novel protein isoforms or may provide an evolutionary benefit for the organism. Additional research is needed to determine the validity of these hypotheses. Clearly, RE are an integral part of the transcriptome and these transcripts should be examined alongside other RNAs. Such studies will provide a more comprehensive annotation of the transcriptome, a better understanding of RE biology, insights into cellular RNA quality control and splicing mechanisms and perhaps may result in the discovery of novel functional gene products.

Materials and Methods

OFC samples

Total RNA from 59 orbitofrontal cortex samples from the Stanley Neuropathology Consortium Collection (11) isolated using the Qiagen RNEasy kit (Qiagen Hilden, Germany) was kindly provided by the Stanley Medical Research Institute. As denoted in Supplementary Material, Table S2, these postmortem samples were collected from 15 subjects with major depression, 15 subjects with bipolar disorder, 14 subjects with schizophrenia and 15 control subjects.

Library preparation and sequencing

Strand-specific RNA-seq libraries were constructed using the TruSeq RNA SamplePrep Guide version 15008136_A with modifications. Briefly, mRNA was purified from 2 μg of total orbitofrontal cortex RNA using Illumina (San Diego, CA) RNA purification beads, the resulting mRNA was fragmented using the Illumina Elute, Prime, Fragment Mix and first strand cDNA was synthesized following the TruSeq RNA protocol. Second strand cDNA was synthesized using 8 μl of 10X NEBNext® Second Strand Synthesis (dNTP-Free) Reaction Buffer, 2μl of 10X SuperScript II RT Buffer (NEB Ipswich, MA), 250μM of each dATP, dUTP, dCTP and dGTP, all of the material from the first strand cDNA reaction and 4μl of second strand enzyme (NEB) in a total of 100μl. The reaction was incubated at 16° C for 2.5 h. The resulting double stranded cDNA was purified, end repaired and adenylated following the TruSeq RNA Sample Prep protocol. One microliter of Illumina adapters were used for the ligation following the TruSeq RNA Sample Prep protocol. The adapter ligated cDNA library was then purified using Ampure beads and subjected to USER enzyme digestion in 5 μl of 10X HotStar PCR buffer (Qiagen Hilden, Germany), 1 unit of USER enzyme (NEB) in a total of 50μl.

This reaction was incubated at 37° C for 15 min and the enzyme was inactivated by heating to 95° C for 5 min. The digested cDNA was purified using Ampure beads and PCR amplified following the protocol in the TruSeq RNA Sample Prep protocol. The 100bp single end sequencing was performed on an Illumina HiSeq 2000 instrument at a depth of roughly 180 million reads per sample.

Human annotation files

We used the UCSC Genome Bioinformatics hg19 RepeatMasker table (http://hgdownload.soe.ucsc.edu/goldenPath/hg19/database/rmsk.txt.gz; date last accessed September 26, 2016) as our annotation file. The RepeatMasker table was created using the Jan 29 2009 (open-3-2-7) version of the RepeatMasker run with the -s (sensitive) setting to screen the hg19 genome build for interspersed repeats and low complexity sequences, based on REEASE 20090120 of library RepeatMaskerLib.embl. In order to remove ambiguity created by overlapping loci in the annotation, we combined overlapping RE into single intervals and treated them as individual repetitive regions. The resulting repetitive annotation file consisted of 5,217,361 non-overlapping repeat regions. As genes can have numerous overlapping exons, we defined a set of nonoverlapping intron and exon regions from the UCSC Genome Bioinformatics hg19 knownGene.txt table. We created one stranded and one non-stranded genomic compartment annotation for the hg19 genome. A base was considered exonic if it was included in any annotated exon regardless of isoform. We subtracted these exons from the known annotated genes to derive our intron list. Regions outside of genes (exons and introns) were designated as intergenic. For the stranded annotation we only combined exons that were located on the same strand and thus, where two genes are encoded in the same region on opposite strands, the stranded annotation contained overlapping exonic regions on each strand. For the non-stranded or strand-free annotation file, a base pair was considered exonic if it was included in any annotated exon on either strand. Thus, this annotation does not contain overlaps between the two strands. We used the stranded annotation for estimating the number of RE in the human orbitofrontal frontal cortex and PBL. The non-stranded or strand-free annotation was used for determining the location of splice junctions. This was necessary because we wanted to compare our results between human data and non-stranded sequencing data obtained from public databases for other species. The non-stranded annotation also ensured that the intergenic and intronic RE that we detect do not result from the exonic sequence mapping to the wrong strand.

The UTR annotation was downloaded from the UCSC Genome Bioinformatics website using the R GenomicFeatures package (27). Repetitive elements were designated to be exonic, intronic, or intergenic if they were completely within a single genic compartment. To determine whether brain expressed repeat element loci (beRE) overlap with lincRNAs, we compared their coordinates to 21,630 lincRNAs and TUCP transcripts (transcripts of uncertain coding potential) annotated in the UCSC Genome Bioinformatics hg19 lincRNAsTranscripts table (16) (http://hgdownload.soe.ucsc.edu/goldenPath/hg19/database/lincRNAsTranscripts.txt.gz; date last accessed September 26, 2016). We also tested whether beRE overlap any of the known and predicted snoRNAs, scaRNAs and microRNAs from snoRNABase (17) and miRBase (18), which are annotated in the UCSC Genome Bioinformatics hg19 wgRna table sno/miRNA track (http://hgdownload.soe.ucsc.edu/goldenPath/hg19/database/wgRna.txt.gz; date last accessed September 26, 2016).

Bowtie 2 alignment

Reads from each sample were aligned to the human genome (hg19) using the short read aligner Bowtie2, which is available at http://bowtie-bio.sourceforge.net/bowtie2/index.shtml (14). Ten bases from the 5 prime and 3 prime ends of the reads were excluded from the genome alignment in order to avoid biases related to random hexamers and sequencing error (28). The '–very-sensitive' option was used for the alignments. Repetitive element expression was estimated using only highly unique reads with a stringent mapping quality score cutoff of 40 or better, which is related to the 'uniqueness' of a read within the genome and indicates that there is a 1/10000 or less chance that the sequencing read is mismapped (14). We counted the number of reads from each sample that surpassed this threshold and aligned to a RepeatMasker annotated RE. This procedure was performed separately for the 'sense' and 'antisense' strands.

Uniqueness estimates

To estimate the number of RE that we can detect using alignments having a mapping quality score of over 40, we counted the number of continuous 75 base intervals that are unique in the human genome in our repeat annotation file. This was accomplished by downloading the 'CRG GEM alignability of 75mers track' (http://hgdownload.soe.ucsc.edu/gbdb/hg19/bbi/wgEncodeCrgMapabilityAlign75mer.bw; date last accessed September 26, 2016). All 75mer intervals that aligned only once to the human genome with up to 2 mismatches were extracted from the CRG track and mapped onto the non-overlapping repetitive loci annotation file. This analysis revealed that over 96% of the non-overlapping annotated repeat loci contain at least one uniquely mapping 75 mer in their sequence based on the uniquely aligning regions from the 'CRG GEM Alignability of 75mers' track. This likely underestimates the mapability of REs both because our reads are longer (81 bases) and because the Bowtie2 aligner (14) is more sensitive than the 2-mismatch algorithm used to produce the alignability track. Regardless, we can estimate expression from the vast majority of annotated repetitive element loci. However, we are unable to measure expression from the younger and potentially active retrotransposons, since there is little or no sequence divergence between these elements.

Unsupervised clustering and RE identification

Unsupervised clustering of the Y-chromosome RE counts (Supplementary Material, Table S3) was performed using the default parameters in the heatmap() function in R (Fig. 1). Loci on the Y-chromosome that had less than 100 reads in all of the samples were excluded from the cluster analysis. In the genome-wide analysis, RE were designated as expressed in the OFC if at least 45 samples (three quarters of the 59 samples) exceeded the threshold for expression (5 or 20 reads). Using the stranded annotation files described above, beRE were mapped to exons, introns, or intergenic regions. Note that RE that overlapped multiple genomic compartments were not included in this analysis.

PBL samples

We performed RNAseq in peripheral blood lymphocytes from 8 individuals. Buffy coats were isolated from whole blood and RNA was isolated using the Zymo (Irvine, CA) Direct-zol RNA

kits. The details of the sequencing for the PBL samples are provided in (Supplementary Material, Table S4). The sequencing libraries were constructed as described for the orbitofrontal cortex samples.

Identifying RE types expressed in the OFC and PBL

RE transcripts expressed in the OFC and PBL were mapped back to the hg19 repeat masker files using the findOverlaps command from the GenomicRanges bioconductor package (27). After subsetting the matching RE loci in the rmsk table, we extracted the 'family', 'category' and 'name' fields. The R command 'table' was run on each field to find the number of RE in each category. RE transcripts that could not be specifically mapped to a single RE locus were excluded from this analysis. A strand free hg19 annotation was used to map RE transcripts to exons, introns and intergenic regions.

Differential RE expression between OFC and peripheral PBL lymphocytes

We compared expression in individuals without a psychiatric diagnosis between the orbitofrontal cortex (n = 15) and the PBL of 8 unrelated individuals. All reads were aligned using Bowtie2 as described in the **Alignment** section. Two count tables were made for the hg19 repeatmasker annotation. One for reads that aligned to the annotation in the sense direction and another for reads that aligned in the antisense direction. DESeq2 with default settings was used to perform a separate analysis for the sense and antisense hits. Hits on the sex chromosomes, those that showed a base mean value of less than 10 reads or had an absolute log fold change of less than one were ignored. To demonstrate that this analysis applied generally to the OFC, we performed a secondary analysis using all 59 OFC samples. Due to the imbalance in sample size, we included all RE with an average of at least 10 reads per sample in the PBL or OFC. Hits on the sex chromosomes or with an absolute log fold change of less than one were ignored, but those with a base mean value of less than 10 were included.

Samples Used for Quantitative PCR

Sample Name	Diagnosis	Sex	Age (years)	Race
PBL-1	Control	Female	22	Caucasian
PBL-2	MDD	Female	60	African American
PBL-3	MDD	Male	25	Caucasian
PBL-4	Control	Male	26	Caucasian
PBL-5	SCZ	Male	63	African American
OFC_Sample-9	MDD	Female	32	Caucasian
OFC_Sample-15	MDD	Male	65	Caucasian
OFC_Sample-18	MDD	Female	56	Caucasian
OFC_Sample-19	SCZ	Male	25	Caucasian
OFC_Sample-29	BPD	Male	30	Caucasian

The human samples were collected in accordance with all institutional regulations and policies. The studies were reviewed and approved by the Sheppard Pratt Institututional Review Board. Blood samples were collected from consenting individuals at Sheppard Pratt Hospital. The OFC samples are described earlier in the methods. The PBL samples were extracted with the Quick-RNA Kits (Zymo Research CAT#: R1054) using the manufacturers recommended protocol. DNAse digestion was performed on column during the RNA isolation process as recommended by the manufacturer. The OFC samples were extracted by SMRI using the RNeasy kit (Qiagen CAT#: 75144). The OFC samples were DNAse digested using the Quick RNA Kits (Zymo Research CAT#: R1054). The integrity of the RNA was confirmed on a tape station instrument (Agilent, DE) and the amount of RNA isolated was measured on a nanodrop 2000 instrument (Thermofisher Scientific, MA).

Quantitative PCR validation of differential expression

Five hundred nanograms of total RNA was converted into cDNA using the SuperScript First-Strand Synthesis kit (Life Technologies, CA CAT#:11904018) following the manufacturers recommended protocol for making random hexamer primed cDNA. The cDNA was diluted 1:10 and 2 μl's were used per PCR reaction. Quantitative PCR reactions (10μl volume) were carried out using the Kapa Probe Fast qPCR kit (Kapa Biosystems, MA CAT#:KK4703) and all reactions were performed in an ABI 7900 instrument. After an initial incubation of 95°C for 3 min, forty cycles of 95°C for 3 seconds and 60°C for 20 s was performed. The primers and probes for the qPCR reactions were ordered from intergrated DNA technologies(IDT IA). The results were normalized to Gapdh (IDT CAT# Hs.PT.39a. 22214836). The Ct value thresholds were selected automatically by the SDS 2.4 software (Life Technologies, CA). One tailed *t*-tests were performed on the Gapdh normalized (delta-Ct) values to calculate p-values.

chr1:157647412 – 157647724

5'-/56-FAM/TCA GAT GGA/ZEN/GTG AGG TAG AGG AAG A/3IABkFQ/-3'

5'-ATC TAT GTG GCT TCT CCC ATT C-3'

5'-AGG AGA GCA TGG GAA TAA TGA A-3'**chr2:231031543 – 231031809**

5'-/56-FAM/TGC TAT AAA/ZEN/TGA GTG ACT GCT GCT TCC T/3IABkFQ/-3'

5'-GAA ACA GGG ACT GAA CAA CTA CTA-3'

5'-CAA CAT GTG CTT GAC TGG ATT T-3'

chr5:63909402 – 63910168

5'-/56-FAM/TCT GAC CCA/ZEN/TTT AAA GGA GAA TTT CCC/3IABkFQ/-3'

5'-ATG CTG CTA TAA ACA TCC TTT CTT G-3'

5'-CCT GCA TGC CTA GTA ATT CCA-3'

chr9:114313981 – 114316038 Set 1

5'-/56-FAM/TGG AGG TTC/ZEN/TTT GAG ACA GAA TAT ACC CAG/3IABkFQ/-3'

5'-GAT ACA AGA CAC TGC CCA AAG A-3'

5'-TGC CAT GAG TTT GAG CCT TTA-3'

chr9:114313981 – 114316038 Set 2

5'-/56-FAM/TGT GCA CTA/ZEN/TAC TGT CTC TAG TCT TCT TGA/3IABkFQ/-3'

5'-GCT GGT ACT ACA ACT GGC ATA G-3'

5'-TTT GCT CTG CTG CAG TAA ATT G-3'

chr17:33595842 – 33596110 Set 1

5'-/56-FAM/ATG AGG TCA/ZEN/CCT CTG TGA CTG CC/3IABkFQ/-3'

5'-CTA ATG CAG GAT GTG GGA AAC A-3'

5'-TGC AGC CTA GAA ACT CTC TAC A-3'

chr17:33595842 – 33596110 Set 2

5'-/56-FAM/ACC TCA TCT/ZEN/GTT TCC CAC ATC CTG C/3IABkFQ/-3'

5'-AGG TCT CGA GAT ATT GGA AAT CAG-3'

5'-AAA CTC TCT ACA GGC AGT AAG C-3'

chr17:38709034-38709442

5'-/56-FAM/TAA TCC CGT/ZEN/TCA CAA GGG TGG AGC/3IABkFQ/-3'

5'-ACT AGT TCC CTC CTG CTC TT-3'

5'-GGG CCT TTG GGA GTT GAT TA-3'

chr19:54944913 – 54945635

5'-/56-FAM/AAG TAG CTT/ZEN/GCC TTC TCT GGG CTG/3IABkFQ/-3'

5'-CTC TTG CAA TGG TGG AAA GAA G-3'

5'-CAC TGG AGC CTA AGA GAA TCA G-3'

Quantitative PCR comparison of RE spliced vs canonical transcript levels

The PCR reactions were carried out as described in the Quantitative PCR Validation of Differential Expression section. Serial dilutions (0.1,0.025,0.00625,0.00015) of brain cDNA constructed from commercially available human brain total RNA (Life Technologies, CA CAT#:AM6050) was used for making standard curves and determining the amplification efficiency of each assay. In order to be consistent identical amounts from the same cDNA synthesis reaction were used to measure both the canonical and the RE splice transcripts quantitative PCR assay in each sample. The SDS 2.4 software was used to automatically determine the Ct threshold. The same 5 OFC samples used for validating differential expression were assayed. A one-tailed, paired *t*-test was used to calculate the reported p-value. All oligonucleotides were ordered from IDT(IA)

GLRB canonical

5'-/56-FAM/CT TCT CAG T/Zen/C CTC AGC TTG GCC TC/3IABkFQ/-3'

5'- CCA AAC ATC AAG AGC CTT CAC -3'

5'- TGG CTT TCC TTC TGG ATC AAC -3'

GLRB repeat variant

5'-/56-FAM/TT CTG CTC T/Zen/T GCC CGA GGT CAT T/3IABkFQ/-3'

5'-/TGT CTC CTA ACT GCT TTC CTT G -3'

5'-/CTC TCC TGG CTT TCC TTC TG -3'

RE splice junction identification

The sequencing reads were aligned using the Tophat2 short read aligner (2.0.8) with the setting –library-type = fr-firststrand and providing the hg19 genes.gtf annotation (20). The resulting junctions.bed file for each sample was used to identify all splice junctions. The annotation file for known genes and repeat masker was downloaded from UCSC and splice junctions that span annotated exons and RE located in intronic or intergenic regions were identified. The genome builds used for the different species are hg19 for human, mm9 for mouse, rn5 for rat, canFam3 for dog, oryCun2 for rabbit, danRer7 for zebrafish and dm3 for *Drosophila melanogaster*.

Differential expression analysis of RE splice junctions in psychiatric disease

DESeq2 with default settings was used to compare the number of sequencing reads overlapping each splice junction in each sample from control individuals to samples from individuals with a psychiatric disorder. A separate analysis was performed to comparing splice site coverage in each psychiatric disorder to the coverage in controls individuals.

Mapping splice junctions to genes and exons

The genes and the exons were obtained from the corresponding transcript database using the 'gene' and 'exons' command in the GenomicFeatures package. The findOverlaps command from GenomicRanges with maxgap parameter set to 10 was used to map the identified donor and acceptor splice sites in exons to the 'gene' and 'exons' list created with the GenomicFeatures package. We used the strand information in the junctions.bed file generated by Tophat 2 for the RE splice junctions and performed a strand specific analysis using the strand information for exons and genes in the respective annotation files. Supplementary Material, Table S13 shows the results of the findOverlaps command when we set the ignore.strand parameter to TRUE. Splice junctions that mapped to more than one gene or exon were only counted once.

Mapping splice junctions to RE elements

The donor and acceptor sites in RE were mapped back to the appropriate repeat masker files using the findOverlaps command from the GenomicRanges bioconductor package. Splice sites that mapped to more than a single RE were only counted once. In identifying RE families, names and categories we excluded splice sites that mapped to multiple RE since we could not assign them to a single RE with certainty.

Mapping to UTRs and coding regions

The UTR and the coding regions were based on the bioconductor TxDb.Hsapiens.UCSC.hg19.knownGene and TxDb.Mmusculus.UCSC.mm9.knownGene for human and mouse respectively. For all other genomes, the command makeTxDbFromUCSC command from the GenomicFeatures package was used with 'refGene' as the parameter (27). We used the splice site (either donor or accepted) in the annotated exon to determine whether the RE spliced into a coding region or UTR. A single splice site may have mapped to multiple annotated transcripts. Thus, we only counted a splice junction once if it always mapped to coding regions or UTRs. However, if a splice junction mapped to a UTR in one transcript and a coding region in another transcript, it was included once for the coding region count and again for the UTR count.

PCR validation of splice junctions

Commercially available human brain total RNA (LifeTech Carlsbad, CA CAT#:AM6050), human peripheral lymphocytes (Clonetech, CA Cat#: 636592) and human genomic DNA (Promega Madison, WI CAT#:G1471) were purchased. One microgram of total RNA was converted into cDNA using the SuperScript First-Strand Synthesis kit (Life Technologies, CA CAT#:11904018). KAPA2G Fast PCR Kits were used for amplification following the manufacturers recommended protocol. The primers were ordered from Integrated DNA Technologies (Corlaville, IA) and had the following sequences:

*AGAP3*_For:AGACCATCGCTGCCTCCT,*AGAP3*_Rev:CTTGGGCAôAACAGCCTTC, *MVD*_For:AGCGGAGGô TTGCAGTGAG,*MVD*_Rev:TGGCGGCôAGTCACô TTGTA,*PQLC1*_For:CAGGACGCAô CTGCACGTA,*PQLC1*_Rev:CTGAô TGAGAAGTCAGCTGTTAATC,*SCAMP5*_For:ACCTGCCATô GATGTTACCAG,*SCAMP5*_Rev:TGCTTCô

AACCTCCCGAGTAG,*GLBR2*_A_For:CACATGCGô
TGGAAGTCATCT,*GLBR2*_A_Rev:CTGCTTTCCTô
TGCTTCTGCT,*GLBR2*_B_For:AGCAôGAAGCAAGGAA
AGCAG,*GLBR2*_B_Rev:CTTTTGG
GGTTGTTCAGCAT,*KCINP4*_A_For:CTCCCTGTôGGAAAô
GAACTGC,*KCINP4*_A_Rev:GTCAAGGGAGAô
GAGCAGGTG,*KCINP4*_B_For:GAGAACAGCAô
TGGGGGAACT,*KCINP4*_B_Rev:AGCGTGGAAGô
ATGAACTGGA,*GPHN*_A_For:AGAAGACCGCô
AGTGGGATAA,*GPHN*_A_Rev:CTTCTCTGCô
CTCTGCCACTC,*GPHN*_B_For:AGGTAGAAGAGô
GCAGGCACA,*GPHN*_B_Rev:CCTTGôATTTCTTô
CTATTTCATCTGG

The primers used for validating the PBL and the intergenic junctions are:

*PPP2R5C*_For:CAAACGGCAAAATô
TCTTCAAGGT,*PPP2R5C*_Rev:CATCACTGAAô
GGATGGCATCA, *FYB*_For: ATCCTCTGTCGGô
GTTGCC,*FYB*_Rev: CAGGAGATCôAGTGAGCTAGCA, *NPR2*_For: GAACAô
TTGTTAGGGACGGCC,*NPR2*_Rev:GTCTTCCCGô
GGCTCTTATCA,*RPL23*_For:AGGATô
TCCAAGTCCAGCAGT,*RPL23*_Rev:GACTTGTGGCô
CCCGGATT,*HSD11B1L*_For:GAGCTGGACGTôGCAGGAC,
*HSD11B1L*_Rev: CTATGGôCTGCTCTCCTCCC

The obtained PCR amplicons were cloned into the pCR2.1 Topo vector from Life Technologies and Sanger sequencing was performed to confirm the splice sites. Note that a small number of the PCR reactions were performed in cDNA and genomic DNA extracted from the same OFC sample to rule out inter-individual differences between RNA and DNA structure.

RE type involved in splicing

We used GenomicRanges to map splice junctions to annotated RE in the appropriate repeat masker (rmsk) file. This analysis was solely based on the splice junction i.e. if an RE contained more than one splice junction, it was counted multiple times. After subsetting the matching RE loci in the rmsk table, we extracted the 'family', 'category' and 'name' fields. The R command 'table' was run on each field to find the number of RE in each category. Splice junctions that could not be specifically mapped to a single RE locus were excluded from this analysis.

Expected vs observed RE types in cortex and PBL

We calculated the proportion of annotated loci of each class in two ways. First, we used the strand-free annotation of the human genome hg19 (described above) and the RE annotations from the UCSC RepeatMasker database to determine the total number of loci of each class of RE in the genome as a whole, in intergenic regions and in introns. The RE in each genomic compartment were identified using the findOverlaps() function with the options type="within", and ignore.strand = TRUE from the GenomicRanges Bioconductor package. After finding the number of loci of each RE class in each genomic compartment, we calculated the relative proportion of loci attributable to each class by dividing the number of loci of each class of RE by the total number of all RE. We calculated the relative proportion of RE base pairs attributable to each class of RE by adding together the lengths of all loci of each class and dividing it by the total lengths of all loci of all classes of RE. The number of splice junction donor and acceptor sites in intergenic regions and introns that would be expected if all bases in all classes of RE were equally likely to form splice junctions was calculated by multiplying the relative proportion of RE base pairs in intergenic regions and introns (described above) by the total number of splice junction donor and acceptor sites in each genomic compartment.

CAGE data comparison

The previously published list of putative alternative human promoters was downloaded (file human_alt_promoter_targets.txt in Supplementary Material, Data 2) (5).This list of 117165 RE loci was mapped onto the hg19 build of the human genome using the liftover tool (https://genome.ucsc.edu/cgi-bin/hgLiftOver; date last accessed September 26, 2016). The resulting coordinates were overlapped with the splice junctions identified in the OFC using the Genomic Ranges package in bioconductor (27) with the maxgap parameter set to 4.

Splice site sequence determination

We used hg19 of the human genome build and extracted the sequences at each intronic RE splice site in the OFC including 3 bases upstream and downstream. The number of each nucleotide at a specific position was tabulated and is presented in Supplementary Material, Table S12.

Sequencing files for non-human organisms

The '.sra' files denoted in Supplementary Material, Table S17 were downloaded from the Sequence Read Archive and processed as described in the **RE splice junction identification** section. The genome builds that were used for the analysis are canFam3 for dog, dm3 for *Drosophila melanogaster*, mm9 for mouse, rn5 for rat, oryCun2 for rabbit, danRer7 for zebrafish. Repeat masker files (rmsk) and knowngene files for each genome were downloaded from the UCSC genome browser.

Polymorphic RE analysis

We downloaded the location of polymorphic RE on hg19 from database of retrotransposon insertion polymorphisms in humans (dbRIP) at http://dbrip.org/searchRIP.html; date last accessed September 26, 2016. The 'in Genome' field was set to hg19 and the 'with insertion identified from' field to UCSC. The locations of the 1593 polymorphic RE were compared with intronic and intergenic splice junctions identified in the orbitofrontal cortex samples.

Coding potential

We used RE that contained an acceptor and a donor splice site. Such splicing events that produced an 'exon' of at least 20 base pairs were included in the analysis. For sequences on the plus strand, we translated the RNA in the +1, +2 and +3 frames, whereas sequences on the minus strand were translated in the -1, -2 and -3 frames. We then counted the number of loci that contained at least one open reading frame.

Coverage

We calculated coverage for RE that contained one 5 prime and one 3 prime splice site. Such splicing events that produced an 'exon' of at least 20 base pairs were included in the analysis.

Coverage is also provided for the annotated upstream and downstream gene exons surrounding the RE. To calculate a single coverage value for each region, we summed the coverage at each base pair and divided it by the length of the region. The values presented represent the average coverage summed for all 59 samples. The accepted_hits.bam files generated by Tophat2 were used for calculating coverage.

RACE

The RACE reactions were performed using the SMARTer RACE 5'/3' Kit from Clonetech (Cat#634858) following the manufacturer's recommended protocol. 1 μg of Human Poly A+ Brain RNA (Clonetech cat#636102) was converted into RACE suitable cDNA using this kit. The gene specific primers used were GAAGGCTGTTTGCCCAAGGCGACA for 3' RACE and CGCCTTGGGCAAACAGCCTTCCG for 5' RACE. Both primers were purchased from IDT. For the 5' RACE reaction 5 cycles of PCR were performed at 94°C for 30 s/72°C 3 min followed by 5 cycles of 94°C for 30 s/70°C 30 s/72°C 3 min and then 25 cycles of 94°C for 30 s/68°C 30 s/72C 3 min. For the 3' RACE reaction 25 cycles of 94°C for 30 s/68°C 30 s/72°C 3 min was performed. The resulting PCR products were cloned and subjected to Sanger sequencing.

TRAP-seq analysis

TRAP seq data files in Supplementary Material, Table S27 were downloaded from the SRA archives. These paired end sequences were aligned onto the mouse mm10 genome using HISAT2 by performing 2 passes. In the initial pass the '–novel-splicesite-infile' option was not used. The output of the '–novel-splicesite-outfile' from the first run was used as the '–novel-splicesite-infile' for the second run. Intronic and intergenic splice variants were identified as described above using the output of the second HISAT2 run. Non-canonical splice junctions that could not be assigned specifically to the plus or the minus strand were excluded from the analysis since donor and acceptor sites could not be identified. The alignment to UTRs and coding regions were based on TxDb.Mmusculus.UCSC.mm10.knownGene. The junctions that did not have a strand assigned were excluded from the exon/gene/Cds/UTR counts but are reported in the supplementary tables.

Supplementary Material

Supplementary Material is available at *HMG* online.

Acknowledgements

The postmortem brain samples were also kindly provided by Stanley Medical Research Institute. We especially want to thank Drs. Maree Webster and E. Fuller Torrey for making available RNA samples and helpful discussions. We are grateful to Dr. Haig Kazazian Jr for his input and critical reading of this manuscript. We thank Faith Dickerson, Emily Severance and Kristin Gressitt for providing us with PBL samples. We also thank Ou Chen, Suad Diglisic, Elizabeth Rubalcaba and Dr. Adam Ewing for technical help.

Conflict of Interest statement. Miranda Darby, Jeffery Leek, Ben Langmead and Sarven Sabunciyan declare no biomedical financial interests or potential conflicts of interest. Robert H. Yolken is a member of the Stanley Medical Research Institute Board of Directors and Scientific Advisory Board. The terms of this arrangement are being managed by the Johns Hopkins University in accordance with its conflict of interest policies.

Funding

This work was supported by a grant from the Stanley Medical Research Institute.

References

1. Huang, C.R., Burns, K.H. and Boeke, J.D. (2012) Active transposition in genomes. *Annu. Rev.Genet.*, **46**, 651–675. 10.1146/annurev-genet-110711-155616.
2. Slotkin, R.K. and Martienssen, R. (2007) Transposable elements and the epigenetic regulation of the genome. *Nat. Rev. Genet.*, **8**, 272–285.
3. Brouha, B., Schustak, J., Badge, R.M., Lutz-Prigge, S., Farley, A.H., Moran, J.V. and Kazazian, H.H. Jr. (2003) Hot L1s account for the bulk of retrotransposition in the human population. *Proc. Natl Acad. Sci. U S A*, **100**, 5280–5285. Epub 2003 Apr 5287.
4. Hancks, D.C. and Kazazian, H.H. Jr. (2012) Active human retrotransposons: variation and disease. *Curr. Opin. Genet. Dev.*, **22**, 191–203. doi: 110.1016/j.gde.2012.1002.1006. Epub 2012 Mar 1018.
5. Faulkner, G.J., Kimura, Y., Daub, C.O., Wani, S., Plessy, C., Irvine, K.M., Schroder, K., Cloonan, N., Steptoe, A.L., Lassmann, T., *et al*. (2009) The regulated retrotransposon transcriptome of mammalian cells. *Nat. Genet.*, **41**, 563–571. doi: 510.1038/ng.1368. Epub 2009 Apr 1019.
6. Wang, T., Zeng, J., Lowe, C.B., Sellers, R.G., Salama, S.R., Yang, M., Burgess, S.M., Brachmann, R.K. and Haussler, D. (2007) Species-specific endogenous retroviruses shape the transcriptional network of the human tumor suppressor protein p53. *Proc. Natl Acad. Sci. U S A*, **104**, 18613–18618. Epub 12007 Nov 18614.
7. Moller-Krull, M., Zemann, A., Roos, C., Brosius, J. and Schmitz, J. (2008) Beyond DNA: RNA editing and steps toward Alu exonization in primates. *J. Mol. Biol.*, **382**, 601–609. doi: 610.1016/j.jmb.2008.1007.1014. Epub 2008 Jul 1016.
8. Zhang, W., Edwards, A., Fan, W., Fang, Z., Deininger, P.,. and Zhang, K. (2013) Inferring the expression variability of human transposable element-derived exons by linear model analysis of deep RNA sequencing data. *BMC Genomics*, **14**, 584.
9. Mi, S., Lee, X., Li, X., Veldman, G.M., Finnerty, H., Racie, L., LaVallie, E., Tang, X.Y., Edouard, P., Howes, S., *et al*. (2000) Syncytin is a captive retroviral envelope protein involved in human placental morphogenesis. *Nature*, **403**, 785–789.
10. Tyekucheva, S., Yolken, R.H., McCombie, W.R., Parla, J., Kramer, M., Wheelan, S.J. and Sabunciyan, S. (2011) Establishing the baseline level of repetitive element expression in the human cortex. *BMC Genomics*, **12**, 495–410. 1186/1471-2164-1112-1495.
11. Torrey, E.F., Webster, M., Knable, M., Johnston, N. and Yolken, R.H. (2000) The stanley foundation brain collection and neuropathology consortium. *Schizophr. Res.*, **44**, 151–155.
12. Mitchell, A.C. and Mirnics, K. (2012) Gene expression profiling of the brain: pondering facts and fiction. *Neurobiol. Dis.*, **45**, 3–7. doi: 10.1016/j.nbd.2011.1006.1001. Epub 2011 Jun 1014.
13. Kumarasinghe, N., Tooney, P.A. and Schall, U. (2012) Finding the needle in the haystack: A review of microarray gene expression research into schizophrenia. *Aust. N. Z. J. Psychiatry*, **46**, 598–610.
14. Langmead, B. and Salzberg, S.L. (2012) Fast gapped-read alignment with Bowtie 2. *Nat. Methods*, **9**, 357–359.

15. NCBI. (2014). NCBI, Vol. 2014. http://www.ncbi.nlm.nih.gov/gene/100133941.
16. Cabili, M.N., Trapnell, C., Goff, L., Koziol, M., Tazon-Vega, B., Regev, A. and Rinn, J.L. (2011) Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. *Genes Dev.*, **25**, 1915–1927.
17. Lestrade, L. and Weber, M.J. (2006) snoRNA-LBME-db, a comprehensive database of human H/ACA and C/D box snoRNAs. *Nucleic Acids Res.*, **34**, D158–D162.
18. Griffiths-Jones, S. (2006) miRBase: the microRNA sequence database. *Methods Mol. Biol.*, **342**, 129–138.
19. Yang, K., Li, J. and Gao, H. (2006) The impact of sample imbalance on identifying differentially expressed genes. *BMC Bioinformatics*, **7**, S8.
20. Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R. and Salzberg, S.L. (2013) TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol.*, **14**, R36. doi: 10.1186/gb-2013-1114-1184-r1136.
21. Wang, J., Song, L., Grover, D., Azrak, S., Batzer, M.A. and Liang, P. (2006) dbRIP: a highly integrated database of retrotransposon insertion polymorphisms in humans. *Hum. Mutat.*, **27**, 323–329.
22. Mellen, M., Ayata, P., Dewell, S., Kriaucionis, S. and Heintz, N. (2012) MeCP2 binds to 5hmC enriched within active genes and accessible chromatin in the nervous system. *Cell*, **151**, 1417–1430. doi: 1410.1016/j.cell.2012.1411.1022.
23. Singer, T., McConnell, M.J., Marchetto, M.C., Coufal, N.G. and Gage, F.H. (2010) LINE-1 retrotransposons: mediators of somatic variation in neuronal genomes? *Trends Neurosci.*, **33**, 345–354. doi: 310.1016/j.tins.2010.1004.1001. Epub 2010 May 1012.
24. Bundo, M., Toyoshima, M., Okada, Y., Akamatsu, W., Ueda, J., Nemoto-Miyauchi, T., Sunaga, F., Toritsuka, M., Ikawa, D., Kakita, A., *et al.* (2014) Increased l1 retrotransposition in the neuronal genome in schizophrenia. *Neuron*, **81**, 306–313.
25. Nagy, E. and Maquat, L.E. (1998) A rule for termination-codon position within intron-containing genes: when nonsense affects RNA abundance. *Trends Biochem. Sci.*, **23**, 198–199.
26. Lykke-Andersen, S. and Jensen, T.H. (2015) Nonsense-mediated mRNA decay: an intricate machinery that shapes transcriptomes. *Nat. Rev. Mol. Cell Biol.*, **16**, 665–677.
27. Lawrence, M., Huber, W., Pages, H., Aboyoun, P., Carlson, M., Gentleman, R., Morgan, M.T. and Carey, V.J. (2013) Software for computing and annotating genomic ranges. *PLoS Comput. Biol.*, **9**, e1003118. doi: 1003110.1001371/journal.pcbi.1003118. Epub 1002013 Aug 1003118.
28. Lin, W., Piskol, R., Tan, M.H.,. and Li, J.B. (2012) Comment on 'Widespread RNA and DNA sequence differences in the human transcriptome'. *Science*, **335**, 1302. author reply 1302.