Total RNA sequencing reveals nascent transcription and widespread co-transcriptional splicing in…

Article *in* Nature Structural & Molecular Biology · November 2011

DOI: 10.1038/nsmb.2143 · Source: PubMed

CITATIONS
121

READS
354

7 authors, including:

Adam Ameur
Uppsala University
80 PUBLICATIONS **4,164** CITATIONS
SEE PROFILE

Ammar Zaghlool
Uppsala University
12 PUBLICATIONS **591** CITATIONS
SEE PROFILE

Lucia Cavelier
Uppsala University
50 PUBLICATIONS **1,212** CITATIONS
SEE PROFILE

Lars Feuk
Uppsala University
129 PUBLICATIONS **18,291** CITATIONS
SEE PROFILE

RESOURCE

nature structural & molecular biology

Total RNA sequencing reveals nascent transcription and widespread co-transcriptional splicing in the human brain

Adam Ameur[1], Ammar Zaghlool[1], Jonatan Halvardson[1], Anna Wetterbom[2], Ulf Gyllensten[1], Lucia Cavelier[1,3] & Lars Feuk[1,3]

Transcriptome sequencing allows for analysis of mature RNAs at base pair resolution. Here we show that RNA-seq can also be used for studying nascent RNAs undergoing transcription. We sequenced total RNA from human brain and liver and found a large fraction of reads (up to 40%) within introns. Intronic RNAs were abundant in brain tissue, particularly for genes involved in axonal growth and synaptic transmission. Moreover, we detected significant differences in intronic RNA levels between fetal and adult brains. We show that the pattern of intronic sequence read coverage is explained by nascent transcription in combination with co-transcriptional splicing. Further analysis of co-transcriptional splicing indicates a correlation between slowly removed introns and alternative splicing. Our data show that sequencing of total RNA provides unique insight into the transcriptional processes in the cell, with particular importance for normal brain development.

RNA sequencing (RNA-seq) has revolutionized transcriptome analysis because of its high throughput, precision and sensitivity[1–3]. Studies of mammalian transcriptomes using RNA-seq show that although most sequence reads are associated with exons in known genes, many are intronic[4–6]. Published RNA-seq data show that intronic sequence read coverage varies among tissues, with high levels found in the human brain[5,6]. The enrichment of intronic RNAs in brain tissue has not been thoroughly investigated, but we find it particularly intriguing given the extreme transcriptome diversity and extensive RNA processing of neuronal cells[7]. Moreover, it has not been demonstrated whether these intronic reads originate from independent transcripts located within introns, or whether they represent immature transcripts that have not yet been spliced. Immature transcripts could comprise either full-length pre-mRNA molecules, or nascent transcripts in which the RNA polymerase has not yet reached the 3′ end of the gene.

Methods that measure nascent transcript formation[8–10] show high levels of RNA across introns but no evidence of independent intronic transcripts. Moreover, nascent transcription profiles reveal that transcription is tightly coupled to splicing[10], a mechanism termed co-transcriptional splicing. Co-transcriptional splicing is the process by which the splicing machinery works behind the RNA polymerase to form spliced products as the polymerase proceeds with transcription[11–13]. Early studies have demonstrated co-transcriptional splicing in single genes in lower eukaryotes and mammalian cells[14–18], and a recent global analysis shows that it occurs in most intron-containing genes in yeast[19]. The extent to which co-transcriptional splicing occurs *in vivo* in mammalian cells is unclear, although studies of specific genes have provided evidence that the mechanism exists[20–22]. Further evidence for co-transcriptional splicing comes from tiling array experiments that showed a typical saw-tooth pattern over long intron-containing genes in tumor necrosis factor alpha (TNF-α)-stimulated human cells[10]. These results demonstrate that transcript profiles can be used to infer co-transcriptional splicing. The identification of the specific transcripts undergoing co-transcriptional splicing is important and would lead to increased understanding of transcriptional regulation, as co-transcriptional splicing implies that the fate of these transcripts is dictated during ongoing transcription.

To determine the pattern of intronic transcription and its association to co-transcriptional splicing, we have sequenced and analyzed expression profiles of the total RNA of human brain and liver tissues. We show that the level of intronic RNA is particularly high in brain compared to liver, and in fetal brain compared to adult brain. Furthermore, we demonstrate that the specific transcript pattern shown by many introns represents nascent transcription combined with co-transcriptional splicing. We show that co-transcriptional splicing is a widespread mechanism, and our results provide unique insights into the *in vivo* transcriptional processes in the developing brain.

RESULTS

Total RNA sequencing reveals immature transcripts

To understand the nature and functional importance of intronic transcription, we compared RNA-seq data from hexamer-primed total RNA to oligo(dT)-primed RNA data from the same sample (frontal cortex of chimpanzee (*Pan troglodytes*))[4]. Similarly to the oligo(dT)-primed data, we found that the majority of mapped sequence reads using hexamer-primed total RNA were located within known genes. However, we observed a higher proportion of RNA originating from regions outside known exons in total RNA (74%) compared to oligo(dT)-primed RNA (20%). Notably, 38% of all mapped sequence

[1]Department of Immunology, Genetics and Pathology, Science for Life Laboratory Uppsala, Rudbeck Laboratory, Uppsala University, Uppsala, Sweden. [2]Science for Life Laboratory, Karolinska Institutet Science Park, Stockholm, Sweden. [3]These authors contributed equally to this work. Correspondence should be addressed to L.C. (lucia.cavelier@igp.uu.se) and L.F. (lars.feuk@igp.uu.se).

Received 18 March; accepted 22 August; published online 6 November 2011; doi:10.1038/nsmb.2143

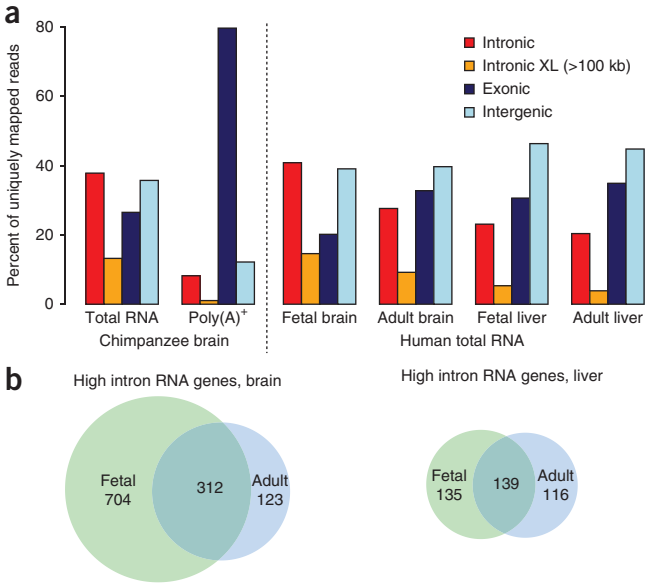

Figure 1 A large proportion of RNA-seq reads map to intronic regions. (**a**) The percentage of reads mapping to intronic, exonic and intergenic regions is shown for each RNA-seq dataset (on the *y* axis). For introns, only reads mapping to the same strand as the surrounding gene are considered. Reads located inside introns but on the opposite strand are labeled intergenic. Introns of length >100 kb (XL introns) are shown by separate bars. (**b**) Venn diagrams containing the number of identified genes with high intronic RNA score in human fetal and adult brain (left) and human fetal and adult liver (right).

reads were located in introns for total RNA, whereas only 8% were located in introns for poly(A)$^+$ RNA (**Fig. 1a**), indicating a large number of immature transcripts. The RNA-seq signals for two genes with high levels of intronic RNA in chimpanzee brain, *GRID2* and *NRXN1*, are shown in **Supplementary Figure 1**, with similar patterns for several other genes.

We then conducted identical RNA-seq experiments on total RNA from human fetal and adult tissues. The human data show a similar distribution of exonic, intronic and intergenic reads to that of chimpanzee (**Fig. 1a**), and, again, we found several examples of genes with high levels of intronic RNA. This pattern has also been seen in published RNA-seq data from total RNA in mouse neuronal ES cells[23], and to some extent in poly(A)$^+$-selected mRNA from the same kind of samples (see **Supplementary Fig. 2**).

In our total RNA data, the fraction of reads mapping to introns was significantly higher in brain compared to liver (**Fig. 1a**); a *Z*-test for two proportions gave *P* values < 10^{-300} for both fetal and adult tissues. Analysis of *de novo* predicted splice junctions showed that about 95% of the detected junctions bridged between two known exon boundaries in the same RefSeq gene, whereas only about 1% connected an exon boundary to a non-exonic region (see **Supplementary Table 1**). This analysis corroborates that intronic reads mainly represent immature transcripts, not processed RNA molecules that have already been spliced, showing that nascent transcripts can be detected by total RNA sequencing and that frontal cortex has substantially higher levels of immature transcripts than does liver.

Genes with high intronic RNA levels

To analyze the differences of intronic RNAs between tissues, we devised an intronic RNA score for each gene. The score was based on *P* values from a Wilcoxon signed-rank test comparing the read coverage across each intron to the experimental background represented by the opposite strand (see **Supplementary Methods**). Each intron was divided into 100 bins of equal size, and the average coverage was computed for the individual bins, ensuring a statistical test based on an equal number of observations regardless of the intron length. In brain tissue we found 1,139 significant genes (Bonferroni-corrected *P* value of <10^{-3}) that passed the selection criteria for high levels of intronic RNA, compared to only 390 such genes in liver tissue. Also, many more genes with high intronic read coverage were found in fetal brain compared to adult brain, whereas there was little difference between fetal and adult liver tissues (**Fig. 1b**).

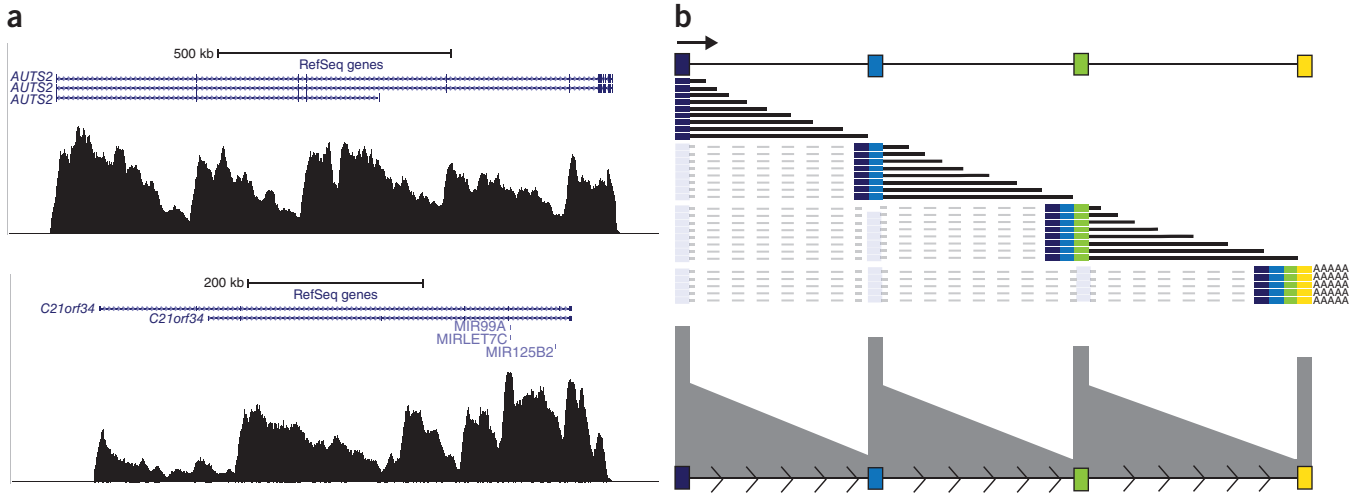

Figure 2 Nascent transcription and co-transcriptional splicing. (**a**) Pattern for *AUTS2* (top) and *C21orf34*, a noncoding RNA gene (bottom), viewed in the University of California, Santa Cruz (UCSC) Genome Browser[47]. The RNA-seq signals have been smoothed using window averaging. For both protein coding genes and long noncoding RNA genes, there is an apparent 'saw-tooth' pattern with higher RNA-seq signal toward the 5′ end of each intron. (**b**) Model for co-transcriptional splicing. The total RNA-seq data give rise to a typical saw-tooth pattern across genes that are actively transcribed. The gradient of RNA across the introns can be explained by a large number of nascent transcripts in various stages of completion. The pattern is repeated for each intron because the nascent transcript is spliced very rapidly after the polymerase completes transcribing each intron. The sequence read coverage is comparatively higher for exons, as the RNA-seq is measuring both the pool of nascent transcripts and the pool of mature polyadenylated RNA.

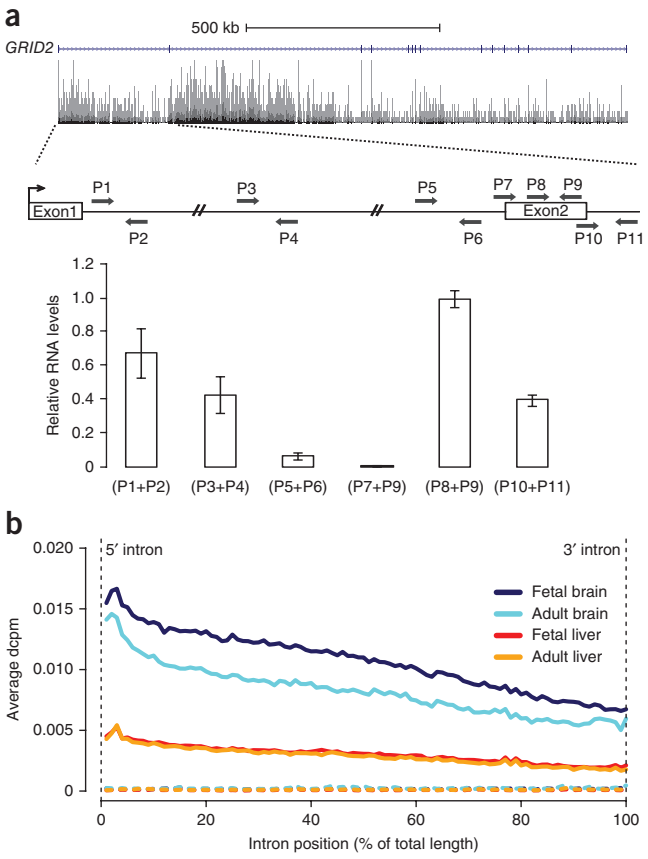

Figure 3 A gradient of RNA levels within introns. (**a**) The relative RNA levels of the first two exons and the surrounding introns of *GRID2* were measured in complementary DNA (cDNA) from human fetal frontal cortex total RNA using quantitative real-time PCR. The qrtPCR results (bottom) correlate with the intronic signals in the RNA-seq data (top). A schematic view of the two first exons of *GRID2* and the intermediate intron is seen in the middle panel. The arrows indicate the location of primers (P1 to P11) used in the experiment. The qrtPCR values are based on three independent experiments (error bars are s.d.). (**b**) RNA-seq signal over all introns in the genome of at least 50 kb in length (L and XL introns), measured in average depth of coverage per million mapped reads (average dcpm). Each intron was divided into 100 bins, and an average value was calculated for each of the individual bins. The four human samples show a decrease of RNA signal across the intron, with the brain samples showing much steeper slopes compared to liver, indicating nascent transcription and co-transcriptional splicing. The dotted lines at the bottom show the RNA-seq signals on the opposite strand.

A list of genes with high intronic RNA levels in all tissues is available (see **Supplementary Data 1**).

We found high levels of nascent transcription for genes in specific functional categories. Gene ontology analysis of the genes in brain tissue highlighted pathways central to neural signaling, with more processes active in fetal than in adult brain and with particular enrichment of cell adhesion, synaptic transmission and neuron projection molecules (**Supplementary Tables 2** and **3**). Pathway analysis further indicated an overrepresentation of genes involved in synaptic signaling pathways in adult and fetal brain and in axonal guidance, ephrin receptor and semaphoring signaling in fetal brain (**Supplementary Fig. 3**). By contrast, genes with high intronic RNA scores in liver were less enriched for specific biological processes, compared to the genes detected in fetal brain (**Supplementary Tables 4** and **5**).

Nascent transcription and co-transcriptional splicing

In addition to the large fraction of intronic reads, specific patterns were evident in the read distribution across introns. For long introns, there was a clear 5′–3′ slope in the read coverage, with substantially higher levels of RNA present in the 5′ end of introns. This pattern was evident for both protein-coding genes and long noncoding RNAs (**Fig. 2a**). Based on these observations, we propose a model to explain the pattern of intronic RNA (**Fig. 2b**). Nascent transcripts at different stages of formation throughout the intron generate a gradient. The fact that this slope spans across individual introns, rather than across the entire transcript, can be explained by co-transcriptional splicing. This model implies that splicing occurs rapidly after transcription is completed for each intron, generating a typical saw-tooth pattern across the transcripts (**Fig. 2b**).

To experimentally validate the transcription patterns seen in our total RNA-seq data, we carried out quantitative real-time PCR (qrtPCR) in two genes with high levels of expression in brain tissue: *NRXN1* and *GRID2*. The results showed an excellent correlation between the qrtPCR and RNA-seq data (**Fig. 3a** and **Supplementary Fig. 4**), confirming the pattern of intronic RNA with high levels at the 5′ end and low levels at the 3′ end of long introns. These experiments show that intronic transcripts are attached to the 5′ exon, whereas there are low levels of fragments connecting the 3′ end of the intron to the downstream exon, as predicted by our model (**Fig. 2b**). To characterize this phenomenon globally, the average read depth in the different samples was plotted for the first and last 500 base pairs (bp) of all introns (**Supplementary Fig. 5**). In the brain, the largest introns (>100 kb in size) had a substantially higher level of intronic RNA in their 5′ ends compared to introns of smaller size. The same pattern was not seen in the liver or in the

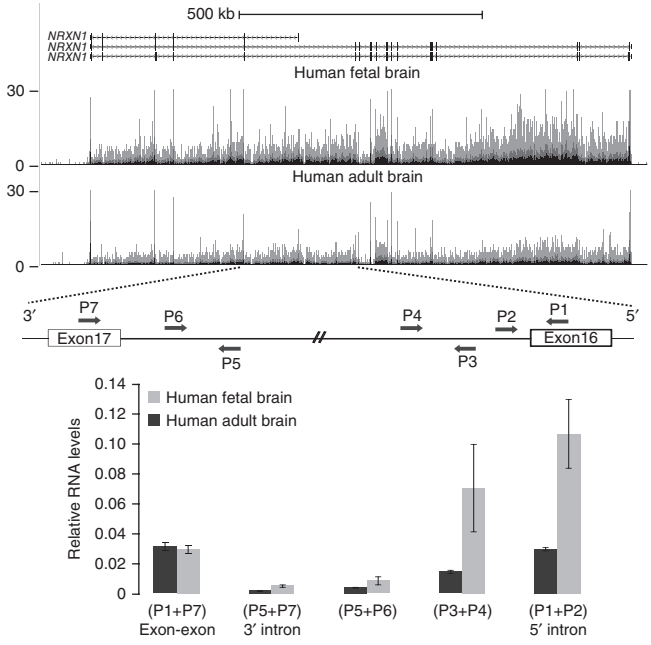

Figure 4 Exonic and intronic RNA levels for *NRXN1*. The relative RNA levels for exons and introns of *NRXN1* were measured by quantitative real-time PCR. The qrtPCR results correlate well with the high intronic signals observed in the RNA-seq data (top). The exonic RNA levels are comparable in fetal and adult frontal cortex, whereas the intronic RNA levels are higher in fetal brain. These results indicate that the level of nascent transcription is not reflected accurately by measuring mature RNA. The qrtPCR values are based on three independent experiments (error bars are s.d.).

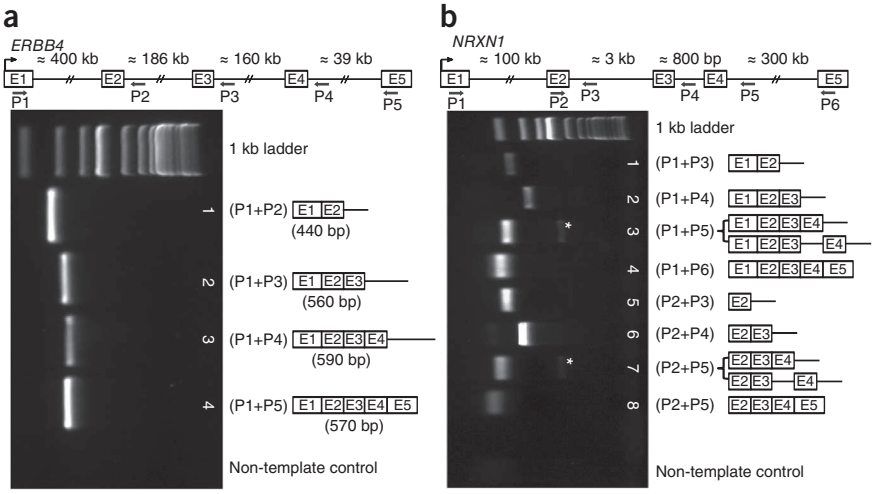

Figure 5 Experimental validation of co-transcriptional splicing in consecutive exons. (**a**) PCR of the *ERBB4* transcripts amplified from fetal frontal cortex cDNA was designed to detect if introns are spliced co-transcriptionally. The schematic representation shows the first five exons (rectangles) and the surrounding introns of *ERBB4*. The arrows indicate the location of primers. Agarose gel electrophoresis of the PCR products (bottom left) reveals that the amplified transcripts are undergoing co-transcriptional splicing. Bottom right, a schematic representation of the intermediate transcripts detected on the gel (validated by Sanger sequencing). (**b**) Similar PCR results for the *NRXN1* in fetal frontal cortex cDNA. Gel bands marked with asterisks (*) represent transcripts where very short introns have not yet been spliced out, providing evidence for a short lag between transcription and splicing.

3′ end of introns. Furthermore, an analysis of all long introns (>50 kb) in the human genome validated this gradient and confirmed the differences that we found between brain and liver, as well as between fetal and adult brain (**Fig. 3b**).

To verify the tissue differences observed in the global analysis, we compared the level of nascent and mature transcripts for *NRXN1* in human fetal and adult frontal cortex using qrtPCR. Although the levels of mature transcripts were equal in the two tissues, the fetal frontal cortex showed relatively higher levels of immature transcripts (**Fig. 4**), indicating a very high rate of transcription combined with a high turnover rate of mature transcripts in the fetal tissue. Alternatively, it is possible that a smaller fraction of immature RNA was processed to mRNA in fetal tissues than in adult tissues, or that pre-mRNA was more stable in fetal than in adult tissues. However, the present data are limited in their capacity to allow us to discriminate between these possibilities. We also note that this high rate of nascent transcription would not be detectable using poly(A)$^+$-selected RNA, implying that total RNA offers additional insight into transcriptional activity and the global RNA profile of a tissue at a given time point.

We also conducted validation experiments demonstrating that consecutive exons of both *ERBB4* and *NRXN1* are co-transcriptionally spliced (see **Fig. 5**), as predicted by our proposed model (**Fig. 2b**). Moreover, our results indicate that although splicing is co-transcriptional, there is a short lag between completion of transcription of an intron and the splicing of that intron (see **Fig. 5b**). These results are in agreement with the previously derived model in which introns are committed to splicing in consecutive order, but the excision may be delayed[24]. To test whether co-transcriptional splicing is restricted to long genes expressed in brain tissue, we conducted similar validation experiments for three additional genes: *TUBB2B*, a short gene expressed in brain (3 kb intron); *ZBTB20*, expressed in liver and brain (>300 kb); and *PAH*, a gene expressed only in liver (80 kb). The investigated introns in all three genes were spliced co-transcriptionally (**Supplementary Fig. 6**), even though some of these co-transcriptional splicing events would be very difficult to detect by analysis of our RNA-seq data because of low read coverage in the intronic regions. In summary, RNA-seq can reveal nascent transcription and co-transcriptional splicing occurring at different levels in human tissues, with the highest levels of intronic expression in fetal brain. However, the RNA-seq data have some inherent limitations and can mainly be used only to study the transcript formation for long introns in genes expressed at relatively high levels.

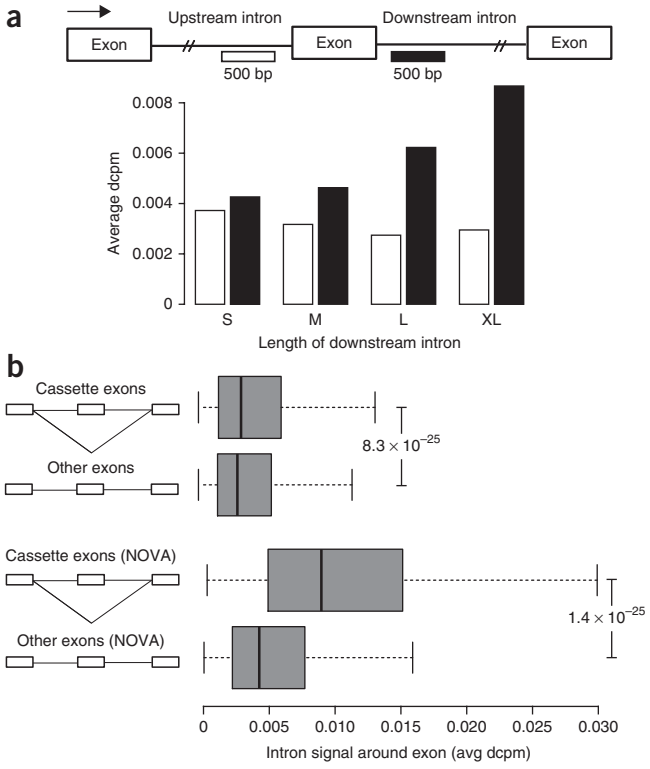

Figure 6 Levels of co-transcriptional splicing within genes. (**a**) In order to quantify co-transcriptional splicing, we measured the read coverage at the 3′ end of the upstream intron (white region) and at the 5′ end of the downstream intron (black region). The evidence for co-transcriptional splicing, represented by the difference in sequence coverage between the two intronic regions, is strongest for the exons succeeded by large introns (>100 kb) and decreases with shorter intron length. S ≤ 10 kb, M = 10–50 kb, L = 50–100 kb, XL ≥ 100 kb. (**b**) Investigation of the correlation between intronic RNA and alternative splicing. On the *x* axis is the intronic RNA level around exons, calculated as the average of reads in the black and white regions outlined in panel **a**. The data show that the read coverage in introns flanking annotated cassette exons is significantly higher than in other introns in the same gene (top two bars), a Wilcoxon test between the two distributions gives a *P* value of 8.3×10^{-25}. A subset of alternatively spliced exons regulated by NOVA show even higher RNA levels in flanking introns (bottom two bars), indicating that these sequences are spliced with different kinetics than other splicing events in the same gene.

Co-transcriptional and alternative splicing

Because the detection of co-transcriptional splicing from RNA-seq data is facilitated by high sequence coverage, we resequenced total RNA from fetal brain tissue, generating approximately 500 million additional reads. We then devised a measure of co-transcriptional splicing for each exon by calculating the difference of intronic read coverage at the 3′ and 5′ ends of the flanking introns (see **Supplementary Methods**). If co-transcriptional splicing occurs as predicted in our model (**Fig. 2b**), the read coverage in the 3′ end of the upstream intron would be expected to be substantially lower than in the 5′ end of the downstream intron. We first correlated these measures with intron length (**Fig. 6a**) and found that evidence for co-transcriptional splicing was strongest for the longest introns, decreasing with shorter intron length. We also estimated that a majority (84%) of exons that were flanked by a large intron showed clear evidence of co-transcriptional splicing (**Supplementary Fig. 7**). We then evaluated whether co-transcriptional splicing is associated with alternative splicing by measuring the read coverage in introns around cassette exons compared to constitutive exons in the same gene. The analysis shows that introns flanking cassette exons have significantly higher RNA levels compared to constitutive exons in the same gene ($P = 8.3 \times 10^{-25}$; **Fig. 6b**). These results are consistent with the idea that alternative exons are spliced at a slower rate than constitutive exons[21,24].

An important question that arises from this analysis is whether alternative splicing affects the intronic expression patterns of genes regulated during neuronal development. Several experimental models have shown that RNA-binding proteins, such as TDP-43 (refs. 25–27) and NOVA[28,29], regulate RNA processing in neuronal cells. We found that genes having high levels of intronic RNA in fetal brain showed a significant overlap with TDP-43 ($P = 6.0 \times 10^{-25}$) and NOVA-2–regulated genes ($P = 8.6 \times 10^{-15}$) (see **Supplementary Methods** and **Supplementary Tables 6–9**). Further analysis showed that the levels of RNA for introns flanking cassette exons regulated by NOVA[29] were significantly higher compared to introns flanking other exons in the same gene ($P = 1.4 \times 10^{-25}$; **Fig. 6b**). These results further support the idea that alternatively spliced exons in neurons are removed at a slower rate than constitutively expressed exons.

DISCUSSION

Recent RNA-seq studies have suggested that unannotated transcripts within introns represent unprocessed transcripts rather than unique independent transcriptional units[30]. Our total RNA-seq data show a high number of reads in intronic regions, especially in large introns of genes expressed in the brain. Our analysis of splice junctions as well as our validation experiments confirm that most intronic RNA indeed stems from unprocessed transcripts. Previous RNA-seq analysis also suggests that poly(A)$^+$ purification is not completely efficient, as a substantial number of intronic reads are retrieved from oligo(dT)-primed RNA[4]. The presence of intronic RNA in poly(A)$^+$ RNA-seq data might represent background oligo(dT) priming to stretches of adenines in primary transcripts, rather than true polyadenylated transcripts. We find further support for this hypothesis from analysis of our oligo(dT)-primed data, in which a higher proportion of intronic reads are flanked by poly(A) stretches (14%) than by poly(T) stretches (9%). A fraction of poly(A)$^+$-selected intronic RNA may also represent transcripts that undergo splicing after polyadenylation.

Intronic RNA levels vary between tissues, with the highest levels found in fetal brain. Independent validation by qrtPCR verified that expression of introns is higher in fetal tissue, whereas exons show similar expression in fetal and adult tissue. We speculate that intronic RNAs in fetal brain are subjected to regulatory pathways specific to the developing brain, where regulatory switches change alternative splice programs to control neuronal development[31,32]. Genes with high levels of intronic RNA in brain are often associated with the synapse, which requires the orchestrated expression of different protein isoforms[7,33] and is highly dependent on RNA processing[25,34,35]. Thus, these processes might require or result in larger intronic RNA levels.

Moreover, several of the genes with high levels of intronic RNA in human frontal cortex have recently been implicated in neurodevelopmental and neuropsychiatric disorders. Of the ten genes with the highest intronic RNA score in fetal brain, four genes (*NRXN1*, *PCDH9*, *MSRA* and *AUTS2*) have been directly implicated in autism, by identification of *de novo* copy number variants (CNVs) or translocations disrupting the gene[36–39]. Large deletions in *MSRA* and *AUTS2* have also been identified in idiopathic epilepsy[40], and *de novo* deletions in *NRXN1* and *AUTS2* have been detected in schizophrenia and people with attention deficit hyperactivity disorder (ADHD)[41–44]. Our top-scoring gene, the noncoding RNA *c21orf34*, showed the strongest association in a previous genome-wide association study with ADHD[45], and the *GAP43* gene has been linked to autism[46]. We therefore suggest that other genes on our list may warrant further investigation in neurodevelopmental disorders.

The levels of intronic RNA were not constant along individual introns. The highest levels were found at the 5′ end of each intron, generating a saw-tooth pattern, which we show is a signature pattern of co-transcriptional splicing. In *Saccharomyces cerevisiae*, co-transcriptional splicing has been shown to occur for the majority of intron-containing genes[19], but in eukaryotes it has only been demonstrated to occur in a handful of genes *in vivo*[10,14,15,21,22]. We also show that the saw-tooth pattern is a general trend, at least for larger introns. It is important to note that our approach is limited in resolution, so we cannot determine from RNA-seq data whether co-transcriptional splicing also occurs for smaller introns. However, we have shown by PCR validation that small introns do undergo co-transcriptional splicing, even when undetected in the sequencing data, and that co-transcriptional splicing occurs also in the liver. Based on our results, we propose that co-transcriptional splicing occurs in the vast majority of long introns in the brain and is also a widespread mechanism in other tissues and for shorter introns.

We observed higher levels of RNA in introns flanking cassette exons, raising the possibility that regulatory factors stabilize these sequences by inhibiting co-transcriptional splicing. This might be particularly important in the developing brain, where differentiation processes require tight regulation of alternative splice variants. Furthermore, we show an overlap between genes regulated by the neuronal RNA-processing proteins TDP-43 and NOVA-2 and genes with high intronic RNA levels in human brain, suggesting that these regulatory proteins might associate to nascent transcripts as they are transcribed. Our results suggest that intronic sequences flanking cassette exons regulated by NOVA proteins are more stable and most probably removed with different kinetics than are other exons in the same gene. This is consistent with the view that alternatively regulated exons show higher levels of intron retention compared to constitutive exons[21].

In summary, we show that sequencing of total RNA gives a very different view of the transcriptional landscape compared to sequencing of poly(A)$^+$ RNA. Our results indicate that co-transcriptional splicing is prevalent in human tissues *in vivo*, and they support a model in which transcription is immediately followed by binding of splicing regulatory proteins to the nascent RNA molecule, which then rapidly undergoes co-transcriptional splicing. We show that the genes that are

most actively transcribed in the human fetal frontal cortex have previously been linked to neurodevelopmental processes, so we propose that further studies of nascent transcription and splicing may provide a better understanding of normal human brain development.

METHODS

Methods and any associated references are available in the online version of the paper at http://www.nature.com/nsmb/.

Accession codes. RNA-seq reads are deposited in the EMBL-EBI Sequence Read Archive (European Nucleotide Archive) under accession number ERP000828.

Note: Supplementary information is available on the Nature Structural & Molecular Biology website.

ACKNOWLEDGMENTS

We thank B. Röken at the Kolmården Zoo for sharing the chimpanzee tissue sample. We also acknowledge the staff members at the Uppsala Genome Center, who conducted the SOLiD sequencing. Financial support for this project was obtained from the Swedish Foundation for Strategic Research (L.F.), the Marcus Borgström Foundation (L.C. and L.F.) and the Göran Gustafsson Foundation (L.F.).

AUTHOR CONTRIBUTIONS

L.C. and L.F. conceived and designed the study. A.A., U.G., L.C. and L.F. coordinated experiments and analysis. A.A., J.H. and A.W. conducted the bioinformatics analysis. A.Z. and L.C. did the sample preparation and experimental analysis. All authors participated in discussions of different parts of the study. A.A., A.Z., L.C. and L.F. wrote the manuscript. All authors read and approved the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Published online at http://www.nature.com/nsmb/.
Reprints and permissions information is available online at http://www.nature.com/reprints/index.html.

1. Marioni, J.C., Mason, C.E., Mane, S.M., Stephens, M. & Gilad, Y. RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome Res.* **18**, 1509–1517 (2008).
2. Mortazavi, A., Williams, B.A., McCue, K., Schaeffer, L. & Wold, B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods* **5**, 621–628 (2008).
3. Sultan, M. *et al.* A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome. *Science* **321**, 956–960 (2008).
4. Wetterbom, A., Ameur, A., Feuk, L., Gyllensten, U. & Cavelier, L. Identification of novel exons and transcribed regions by chimpanzee transcriptome sequencing. *Genome Biol.* **11**, R78 (2010).
5. van Bakel, H., Nislow, C., Blencowe, B.J. & Hughes, T.R. Most "dark matter" transcripts are associated with known genes. *PLoS Biol.* **8**, e1000371 (2010).
6. Kapranov, P. *et al.* The majority of total nuclear-encoded non-ribosomal RNA in a human cell is 'dark matter' un-annotated RNA. *BMC Biol.* **8**, 149 (2010).
7. Blencowe, B.J. Splicing on the brain. *Nat. Genet.* **37**, 796–797 (2005).
8. Churchman, L.S. & Weissman, J.S. Nascent transcript sequencing visualizes transcription at nucleotide resolution. *Nature* **469**, 368–373 (2011).
9. Core, L.J., Waterfall, J.J. & Lis, J.T. Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. *Science* **322**, 1845–1848 (2008).
10. Wada, Y. *et al.* A wave of nascent transcription on activated human genes. *Proc. Natl. Acad. Sci. USA* **106**, 18357–18361 (2009).
11. Bentley, D.L. Rules of engagement: co-transcriptional recruitment of pre-mRNA processing factors. *Curr. Opin. Cell Biol.* **17**, 251–256 (2005).
12. Goldstrohm, A.C., Greenleaf, A.L. & Garcia-Blanco, M.A. Co-transcriptional splicing of pre-messenger RNAs: considerations for the mechanism of alternative splicing. *Gene* **277**, 31–47 (2001).
13. Kornblihtt, A.R., de la Mata, M., Fededa, J.P., Munoz, M.J. & Nogues, G. Multiple links between transcription and splicing. *RNA* **10**, 1489–1498 (2004).
14. Baurén, G. & Wieslander, L. Splicing of Balbiani ring 1 gene pre-mRNA occurs simultaneously with transcription. *Cell* **76**, 183–192 (1994).
15. Beyer, A.L. & Osheim, Y.N. Splice site selection, rate of splicing, and alternative splicing on nascent transcripts. *Genes Dev.* **2**, 754–765 (1988).
16. Kiseleva, E., Wurtz, T., Visa, N. & Daneholt, B. Assembly and disassembly of spliceosomes along a specific pre-messenger RNP fiber. *EMBO J.* **13**, 6052–6061 (1994).
17. Osheim, Y.N., Miller, O.L. & Beyer, A.L. RNP particles at splice junction sequences on *Drosophila* chorion transcripts. *Cell* **43**, 143–151 (1985).
18. Wetterberg, I., Bauren, G. & Wieslander, L. The intranuclear site of excision of each intron in Balbiani ring 3 pre-mRNA is influenced by the time remaining to transcription termination and different excision efficiencies for the various introns. *RNA* **2**, 641–651 (1996).
19. Carrillo Oesterreich, F., Preibisch, S. & Neugebauer, K.M. Global analysis of nascent RNA reveals transcriptional pausing in terminal exons. *Mol. Cell* **40**, 571–581 (2010).
20. Listerman, I., Sapra, A.K. & Neugebauer, K.M. Cotranscriptional coupling of splicing factor recruitment and precursor messenger RNA splicing in mammalian cells. *Nat. Struct. Mol. Biol.* **13**, 815–822 (2006).
21. Pandya-Jones, A. & Black, D.L. Co-transcriptional splicing of constitutive and alternative exons. *RNA* **15**, 1896–1908 (2009).
22. Singh, J. & Padgett, R.A. Rates of *in situ* transcription and splicing in large human genes. *Nat. Struct. Mol. Biol.* **16**, 1128–1133 (2009).
23. Kim, T.K. *et al.* Widespread transcription at neuronal activity-regulated enhancers. *Nature* **465**, 182–187 (2010).
24. de la Mata, M., Lafaille, C. & Kornblihtt, A.R. First come, first served revisited: factors affecting the same alternative splicing event have different effects on the relative rates of intron removal. *RNA* **16**, 904–912 (2010).
25. Polymenidou, M. *et al.* Long pre-mRNA depletion and RNA missplicing contribute to neuronal vulnerability from loss of TDP-43. *Nat. Neurosci.* **14**, 459–468 (2011).
26. Sephton, C.F. *et al.* Identification of neuronal RNA targets of TDP-43-containing ribonucleoprotein complexes. *J. Biol. Chem.* **286**, 1204–1215 (2011).
27. Tollervey, J.R. *et al.* Characterizing the RNA targets and position-dependent splicing regulation by TDP-43. *Nat. Neurosci.* **14**, 452–458 (2011).
28. Ule, J. *et al.* Nova regulates brain-specific splicing to shape the synapse. *Nat. Genet.* **37**, 844–852 (2005).
29. Zhang, C. *et al.* Integrative modeling defines the Nova splicing-regulatory network and its combinatorial controls. *Science* **329**, 439–443 (2010).
30. van Bakel, H., Nislow, C., Blencowe, B.J. & Hughes, T.R. Response to "the reality of pervasive transcription". *PLoS Biol.* **9**, e1001102 (2011).
31. Boutz, P.L. *et al.* A post-transcriptional regulatory switch in polypyrimidine tract-binding proteins reprograms alternative splicing in developing neurons. *Genes Dev.* **21**, 1636–1652 (2007).
32. Grabowski, P.J. RNA-binding proteins switch gears to drive alternative splicing in neurons. *Nat. Struct. Mol. Biol.* **14**, 577–579 (2007).
33. Li, Q., Lee, J.A. & Black, D.L. Neuronal regulation of alternative pre-mRNA splicing. *Nat. Rev. Neurosci.* **8**, 819–831 (2007).
34. Rabin, S.J. *et al.* Sporadic ALS has compartment-specific aberrant exon splicing and altered cell-matrix adhesion biology. *Hum. Mol. Genet.* **19**, 313–328 (2010).
35. Dredge, B.K., Polydorides, A.D. & Darnell, R.B. The splice of life: alternative splicing and neurological disease. *Nat. Rev. Neurosci.* **2**, 43–50 (2001).
36. Kim, H.G. *et al.* Disruption of neurexin 1 associated with autism spectrum disorder. *Am. J. Hum. Genet.* **82**, 199–207 (2008).
37. Marshall, C.R. *et al.* Structural variation of chromosomes in autism spectrum disorder. *Am. J. Hum. Genet.* **82**, 477–488 (2008).
38. Walss-Bass, C. *et al.* Methionine sulfoxide reductase: a novel schizophrenia candidate gene. *Am. J. Med. Genet. B. Neuropsychiatr. Genet.* **150B**, 219–225 (2009).
39. Kalscheuer, V.M. *et al.* Mutations in autism susceptibility candidate 2 (AUTS2) in patients with mental retardation. *Hum. Genet.* **121**, 501–509 (2007).
40. Mefford, H.C. *et al.* Genome-wide copy number variation in epilepsy: novel susceptibility loci in idiopathic generalized and focal epilepsies. *PLoS Genet.* **6**, e1000962 (2010).
41. Rujescu, D. *et al.* Disruption of the neurexin 1 gene is associated with schizophrenia. *Hum. Mol. Genet.* **18**, 988–996 (2009).
42. Kirov, G. *et al.* Comparative genome hybridization suggests a role for NRXN1 and APBA2 in schizophrenia. *Hum. Mol. Genet.* **17**, 458–465 (2008).
43. Vrijenhoek, T. *et al.* Recurrent CNVs disrupt three candidate genes in schizophrenia patients. *Am. J. Hum. Genet.* **83**, 504–510 (2008).
44. Elia, J. *et al.* Rare structural variants found in attention-deficit hyperactivity disorder are preferentially associated with neurodevelopmental genes. *Mol. Psychiatry* **15**, 637–646 (2010).
45. Mick, E. *et al.* Family-based genome-wide association scan of attention-deficit/hyperactivity disorder. *J. Am. Acad. Child Adolesc. Psychiatry* **49**, 898–905.e3 (2010).
46. Zaccaria, K.J., Lagace, D.C., Eisch, A.J. & McCasland, J.S. Resistance to change and vulnerability to stress: autistic-like features of GAP43-deficient mice. *Genes Brain Behav.* **9**, 985–996 (2010).
47. Kent, W.J. *et al.* The human genome browser at UCSC. *Genome Res.* **12**, 996–1006 (2002).

ONLINE METHODS

Preparation of total RNA samples. A tissue sample from chimpanzee frontal cortex was obtained through autopsy of a young chimpanzee from Kolmården Zoo, Sweden. The deep frozen tissue was cryosectioned and the slices recovered in 1 ml of TRIReagent (Applied Biosystems) solution in 1.5 ml tubes. RNA was prepared from three tubes each containing fifteen 10-µm frontal cortex slices. The RiboPure kit (Applied Biosystems) for small tissue samples was used to prepare total RNA, which was subsequently treated with TURBO DNase (Applied Biosystems) to remove possible contamination from genomic DNA, using manufacturer instructions. The amount of RNA was monitored using a nanodrop spectrophotometer. Human total RNA was purchased from BioChain. Total RNA samples were depleted from rRNA using the RiboMinus Eukaryote Kit for RNA and the RiboMinus Concentration Module. Total RNA was used to prepare a fragment library using the SOLiD Total RNA-seq Kit (Applied Biosystems).

SOLiD sequencing and processing of reads. The libraries were sequenced using the AB SOLiD4 system. The read length was 50 bp for all samples and directionality of RNA molecules was preserved in the sequencing. Reads were aligned using version 1.1 of the Applied Biosystems whole transcriptome analysis tool (http://solidsoftwaretools.com/gf/project/transcriptome/). The panTro2 reference sequence was used for the chimpanzee sample and the human genome hg19 assembly version was used for the human samples. For each sample, all reads that mapped to identical positions on the same strand were merged and only counted once, thereby reducing potential experimental biases caused by uneven PCR amplification of transcripts. From these merged reads we constructed a coverage signal over the entire reference sequence. SICTIN[48] was used to enable fast access to the RNA-seq read coverage signal for any selected region of the genome.

Coordinates for genes, introns and alternatively spliced exons. Gene coordinates were downloaded from the RefGene tables in the UCSC Genome Browser[47]. Coordinates for exons and introns were extracted for each individual RefSeq transcript. Alternatively spliced exons were extracted from the Alt Events table in the UCSC Genome Browser[47]. This table contains coordinates for human cassette exons, regardless of the tissue in which they are expressed. To obtain a specific list of cassette exons expressed in the brain, we extracted coordinates for NOVA-regulated exons in mouse brain[29]. The mouse coordinates were translated to hg19 using the UCSC LiftOver tool, and each exon was compared to the list of cassette exons previously obtained. This resulted in a list of 234 exons that were NOVA-regulated in mouse brain tissue, that were also reported as cassette exons.

Statistical analyses of RNA-seq signals. To detect genes with high levels of RNA, we calculated a score for intronic RNA for every RefSeq intron. A region classed as an intron for one RefSeq transcript may contain an exon for some other transcript. To ensure that a highly expressed exon within an intron would only have a minimal effect on the intron RNA score, we designed a statistical method to specifically detect introns with RNA-seq reads distributed across the whole length of the intron in the same orientation as the surrounding gene. In brief, we divided each intron into 100 bins of equal size and conducted a Wilcoxon signed-rank test comparing the average read coverage in each bin against the average read coverage on the antisense strand.

The analysis of co-transcriptional splicing was carried out on RNA-seq data for the human fetal brain, which was sequenced at high coverage (492 million reads). For each internal exon of each RefSeq transcript, we extracted the RNA-seq signals in one window downstream of the exon and one window of equal length upstream of the same exon (see **Fig. 6a**). By calculating the difference in average read coverage between the two windows, we obtained a measure of the level of co-transcriptional splicing for all internal exons. For more details on the statistical analyses, see **Supplementary Methods**.

Gene ontology analysis. The DAVID functional annotation tool[49] was used to conduct gene ontology classification of various lists of RefSeq genes. We generally used all genes in the genome as a background and considered a corrected (Benjamini) P value < 0.01 to be significant. The canonical pathway analysis was conducted using the Ingenuity Pathways Analysis (Ingenuity Systems, http://www.ingenuity.com/).

Splice junction detection. *De novo* splice junctions were detected using version 1.3.2 of SplitSeek[50]. We required each junction to be supported by at least two uniquely mapping reads and to span a distance of at most 2 Mb.

Quantitative real-time PCR. Quantitative real-time PCR was used to validate intronic RNA levels in chimpanzee frontal cortex, adult human frontal cortex and fetal human frontal cortex tissues. The qrtPCR was carried out with Stratagene's Mx3000P, in 96-well plates. A total of 25 µl of reaction solution contained 12.5 µl Maxima SYBR Green, ROX qrtPCR Master Mix (Fermentas), 2.5 ng of cDNA and 200 nM of each primer (**Supplementary Table 10**) in a final volume of 25 µl. The reactions were initiated with 10 min of enzyme activation at 95 °C followed by 40 cycles of denaturation at 95 °C for 15 s, primer annealing at 58 °C for 30 s and extension at 72 °C for 30 s. All samples were amplified in triplicates, and the mean value was used for the final calculations. The target mRNA in the samples was measured based on the corresponding standard curve made for each primer pair. Expression levels of target mRNA were normalized to the levels of β-actin in the same sample. Raw data were analyzed using MxPro (Stratagene).

Validation of co-transcriptional splicing by PCR. To investigate how the introns in *ERBB4*, *NRXN1*, *TUBB2B*, *ZBTB20* and *PAH* are spliced in brain and liver, we carried out a PCR with initial denaturation at 95 °C for 1 min followed by 35 cycles of 95 °C for 15 s, 60 °C for 30 s and 72 °C for 3 min. The reaction contained 10X Advantage PCR buffer (Clontech), 200 µM of each of the deoxyribonucleotide triphosphates (dNTPs), 1 U of Advantage 2 polymerase mix (Clontech), primers (**Supplementary Table 10**) and 2.5 ng of cDNA.

48. Enroth, S., Andersson, R., Wadelius, C. & Komorowski, J. SICTIN: Rapid footprinting of massively parallel sequencing data. *BioData Min.* **3**, 4.
49. Huang da, W., Sherman, B.T. & Lempicki, R.A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.* **4**, 44–57 (2009).
50. Ameur, A., Wetterbom, A., Feuk, L. & Gyllensten, U. Global and unbiased detection of splice junctions from RNA-seq data. *Genome Biol.* **11**, R34 (2010).