Deep sequencing of subcellular RNA fractions shows splicing to be predominantly co-transcriptional in the human genome but inefficient for lncRNAs

Hagen Tilgner,[1,3] David G. Knowles,[1] Rory Johnson,[1] Carrie A. Davis,[2] Sudipto Chakrabortty,[2] Sarah Djebali,[1] João Curado,[1] Michael Snyder,[3] Thomas R. Gingeras,[2] and Roderic Guigó[1,4]

[1]*Centre for Genomic Regulation (CRG) and UPF, E-08003, Barcelona, Catalonia, Spain;* [2]*Cold Spring Harbor Laboratory, Cold Spring Harbor, New York 11724, USA;* [3]*Department of Genetics, Stanford University School of Medicine, Stanford, California 94305, USA*

Splicing remains an incompletely understood process. Recent findings suggest that chromatin structure participates in its regulation. Here, we analyze the RNA from subcellular fractions obtained through RNA-seq in the cell line K562. We show that in the human genome, splicing occurs predominantly during transcription. We introduce the coSI measure, based on RNA-seq reads mapping to exon junctions and borders, to assess the degree of splicing completion around internal exons. We show that, as expected, splicing is almost fully completed in cytosolic polyA+ RNA. In chromatin-associated RNA (which includes the RNA that is being transcribed), for 5.6% of exons, the removal of the surrounding introns is fully completed, compared with 0.3% of exons for which no intron-removal has occurred. The remaining exons exist as a mixture of spliced and fewer unspliced molecules, with a median coSI of 0.75. Thus, most RNAs undergo splicing while being transcribed: "co-transcriptional splicing." Consistent with co-transcriptional spliceosome assembly and splicing, we have found significant enrichment of spliceosomal snRNAs in chromatin-associated RNA compared with other cellular RNA fractions and other nonspliceosomal snRNAs. CoSI scores decrease along the gene, pointing to a "first transcribed, first spliced" rule, yet more downstream exons carry other characteristics, favoring rapid, co-transcriptional intron removal. Exons with low coSI values, that is, in the process of being spliced, are enriched with chromatin marks, consistent with a role for chromatin in splicing during transcription. For alternative exons and long noncoding RNAs, splicing tends to occur later, and the latter might remain unspliced in some cases.

[Supplemental material is available for this article.]

Central in the pathway leading from primary transcripts to mature functional RNAs is splicing, the process by which intervening sequences in the primary transcript (introns) are excised and the remaining sequences (exons) are concatenated together to form the mature eukaryotic RNAs. Conserved sequence motifs, the splice sites, mark exon–intron boundaries and are recognized by elements of the splicing machinery. Splice site sequences, however, do not carry enough information to unequivocally specify exon–intron boundaries, and a plethora of other sequence motifs, recognized by a variety of RNA binding proteins, contribute to define and regulate splice site selection (Graveley 2000; Smith and Valcárcel 2000; Wang and Burge 2008). While there have been considerable advances in modeling splicing from features in the primary transcript sequence (Wang et al. 2004; Barash et al. 2010), it is currently close to impossible to predict from the analysis of mammalian primary RNA sequence alone neither the entire exon–intron structure of transcripts nor their tissue specific expression pattern (i.e., the abundance of given transcript in a given cell type).

It appears thus that other factors, not necessarily encoded in the sequence of the primary transcript, may play a role in splicing definition. Indeed, there is a growing body of evidence suggesting that chromatin structure could play a role in splicing. A number of reports have demonstrated that eukaryotic exonic sequences are enriched in positioned nucleosomes and that some histone modifications show characteristic exonic patterns (Andersson et al. 2009; Hon et al. 2009; Kolasinska-Zwierz et al. 2009; Nahkuri et al. 2009; Schwartz et al. 2009; Spies et al. 2009; Tilgner et al. 2009). Intragenic histone modifications and chromatin structure influences on alternative splicing events have been documented in detail for the fibronectin and *FGFR2* gene (Allo et al. 2009; Schor et al. 2009; Luco et al. 2010), and CTCF-mediated local RNA polymerase II (Pol II) pausing has been shown to influence alternative splicing (Shukla et al. 2011). While the underlying molecular mechanisms connecting chromatin structure with splicing are largely unknown, they require, in principle, for splicing to be somehow connected to transcription. That splicing can be carried out during transcription has been known for a long time (Beyer and Osheim 1988), and increasing evidence exists of coupling between transcription and splicing (Cramer et al. 1997; Roberts et al. 1998; Kadener et al. 2001; Nogues et al. 2002; de la Mata et al. 2003; Howe et al. 2003). Intron removal during transcription has been shown to be predominant in the intron-poor genome of *Saccharomyces cerevisiae* (Carrillo Oesterreich et al. 2010), and recently, Ameur et al. (2011) have proposed co-transcriptional splicing to be widespread in the human brain, based on the analysis of whole-cell, total RNA sequencing (RNA-seq). Here, we analyze the RNA that is still residing on the chromatin template,

[4]Corresponding author
E-mail roderic.guigo@crg.cat
Article and supplemental material are at http://www.genome.org/cgi/doi/10.1101/gr.134445.111. Freely available online through the *Genome Research* Open Access option.

22:1616–1625 © 2012, Published by Cold Spring Harbor Laboratory Press; ISSN 1088-9051/12; www.genome.org

as well as nuclear and cytosolic RNA in its polyadenylated and nonpolyadenylated form. This fractional approach, including separate RNA-seq in the cytosol, enables us to define splicing completion at multiple nuclear stages. Thus we can define the fraction of splicing events around an exon that is co-transcriptional. This, in turn allows the definition of smaller subsets of candidates "with a tendency for post-transcriptional splicing"(postTS) without being biased by intron-retention events. We show that co-transcriptional splicing is predominant in the human genome, providing the basis for the understanding of the role of chromatin structure in splicing definition and regulation. We investigate the rules that determine whether the introns around an exon are to be spliced co-transcriptionally. We also observe a 5′-to-3′ trend in splicing completion, which causes more downstream splicing events to be more prone to postTS and makes the distance to the polyA-site one of the most important factors in determining when splicing is completed. This 5′-to-3′ trend is countered by shorter introns and stronger splice site strengths toward the end of the gene—two features that we have found to promote early, co-transcriptional splicing. Further significant predictors of co-transcriptional splicing include the rate of gene transcription and covalent histone modifications. Long noncoding RNAs (lncRNAs) appear to be less efficiently spliced than protein coding genes and, on occasion, may even remain unspliced. Exons, for which the surrounding introns are in the process of being spliced, are enriched with chromatin marks, consistent with a role for chromatin in splicing during transcription, and splicing around alternative exons is, on average, more post-transcriptional than for constitutive exons.

Results

Deep RNA-seq of subcellular fractions

We have used deep RNA-seq, performed within the framework of the ENCODE project, to interrogate with unprecedented resolution distinct RNA fractions from a number of cellular compartments in human immortalized myelogenous leukemia cells K562: chromatin-associated total RNA, polyA− and polyA+ nuclear RNA, as well as polyA− and polyA+ cytosolic RNA (Fig. 1A; see Djebali et al. 2012 and Supplemental Information/Methods for details on RNA fractionating and controls, sequencing, and bioinformatic analysis of the sequence data). Monitoring of these compartments provides snapshots of the different stages of RNA processing within the cell.

The completed splicing index

We introduce a measure, based on the RNA-seq reads mapping to the exon junctions, to assess the degree of completion of splicing around internal exons. We simply count the number of reads mapping across the exon boundaries into the adjacent intron sequence (which originate from primary, unspliced mRNA molecules), as well as the number of reads split-mapping across exon–exon junctions, either from the exon to another exon of the same gene or between an upstream and a downstream exon (both types of read originating from a successfully completed splicing event) (Fig. 1B). Based on these numbers, we compute the completed splicing index (coSI) of a given exon, corresponding thereby to the weighted percentage of reads supporting splicing completion

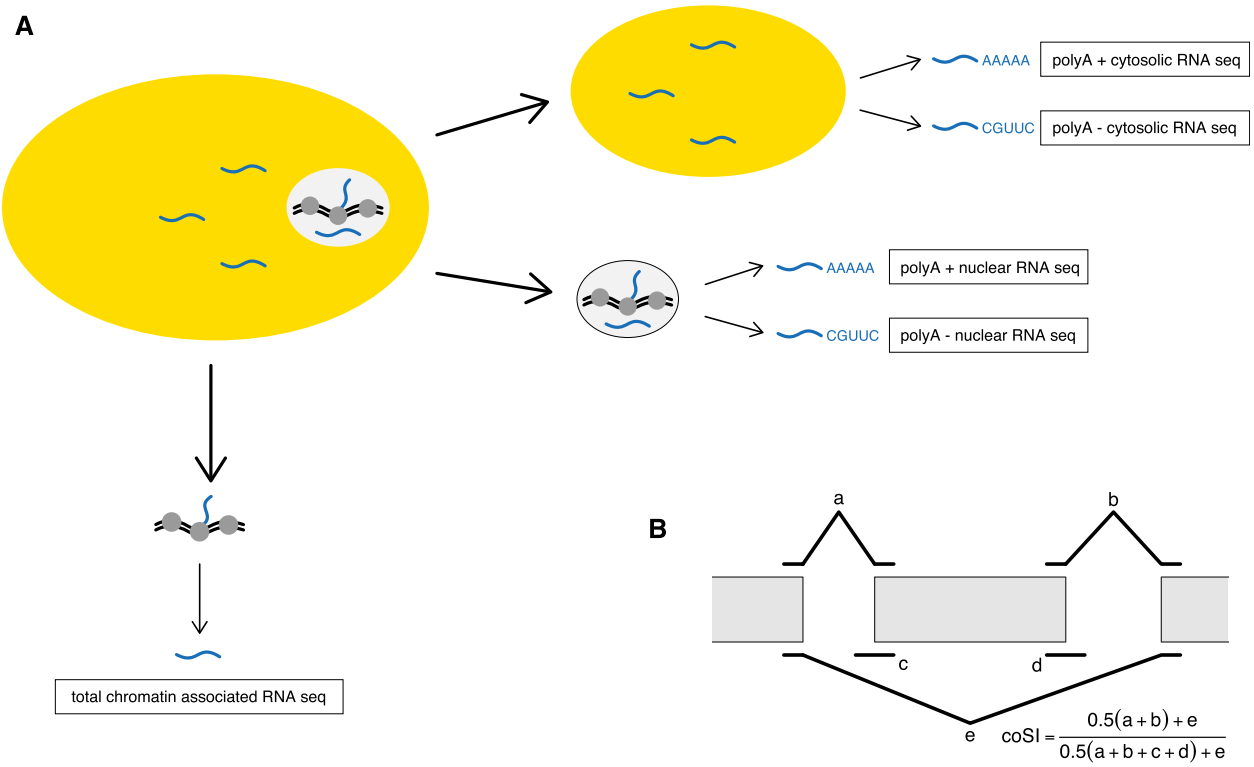

Figure 1. (*A*) Long RNA-seq data sets used in this analysis. (*B*) Definition of the completed splicing index (coSI) for each internal exon and each RNA-seq data set.

around the exon. The coSI value can be broadly assumed to correspond to the fraction of exon-containing RNA molecules in which splicing in the region around the exon has already been carried out. A coSI value of 1 means entirely completed splicing, while coSI = 0 indicates that the exon is still completely included in the sequence of the primary transcript.

Most human exons are already partially spliced in chromatin-associated RNA

We have computed coSI scores for human internal exons (see Supplemental Methods) in all analyzed K562 RNA fractions (Supplemental Table S1). We have observed a higher correlation of coSI values (R = 0.82) between two replicates of the chromatin fraction than between the chromatin fraction and other fractions (R of between 0.35 and 0.71) (Supplemental Fig. S1A–G), confirming that overall coSI values are reproducible within a given experiment.

Figure 2 shows the distribution of coSI scores in the different RNA fractions that we have interrogated. As expected, for most exons, splicing of the corresponding introns is fully completed in the cytosolic polyA+ fraction (92% of the exons have a coSI ≥ 0.95), as well as the cytosolic polyA− fraction (data not shown). Of even more interest with respect to splicing is the polyA− nuclear fraction, in which the median coSI is 0.84. For 16% of the exons, their surrounding introns are completely spliced in this fraction, and only for a vanishing fraction (<0.2% with coSI ≤ 0.05) do the corresponding introns remain completely unspliced. The polyA− nuclear fraction contains RNA molecules of three types: first, RNAs that are still being transcribed and for which transcription has not yet reached the polyA-site; second, RNAs that have been released from the transcribing Pol II, before it could reach the polyA-site; and, third, products of aborted transcription. The high degree of splicing completion in this fraction therefore suggests that splicing is mostly initiated before completion of transcription. Even more enriched for RNAs in the act of being transcribed is the chromatin-associated fraction. With a median coSI of 0.75 in this fraction, around most exons we see large amounts of completed splicing. For 5.6% of the exons, we see absolutely completely spliced introns (coSI ≥ 0.95); however, as in the polyA− nuclear fraction, only a tiny fraction of exons (<0.3%, coSI ≤ 0.05) are surrounded

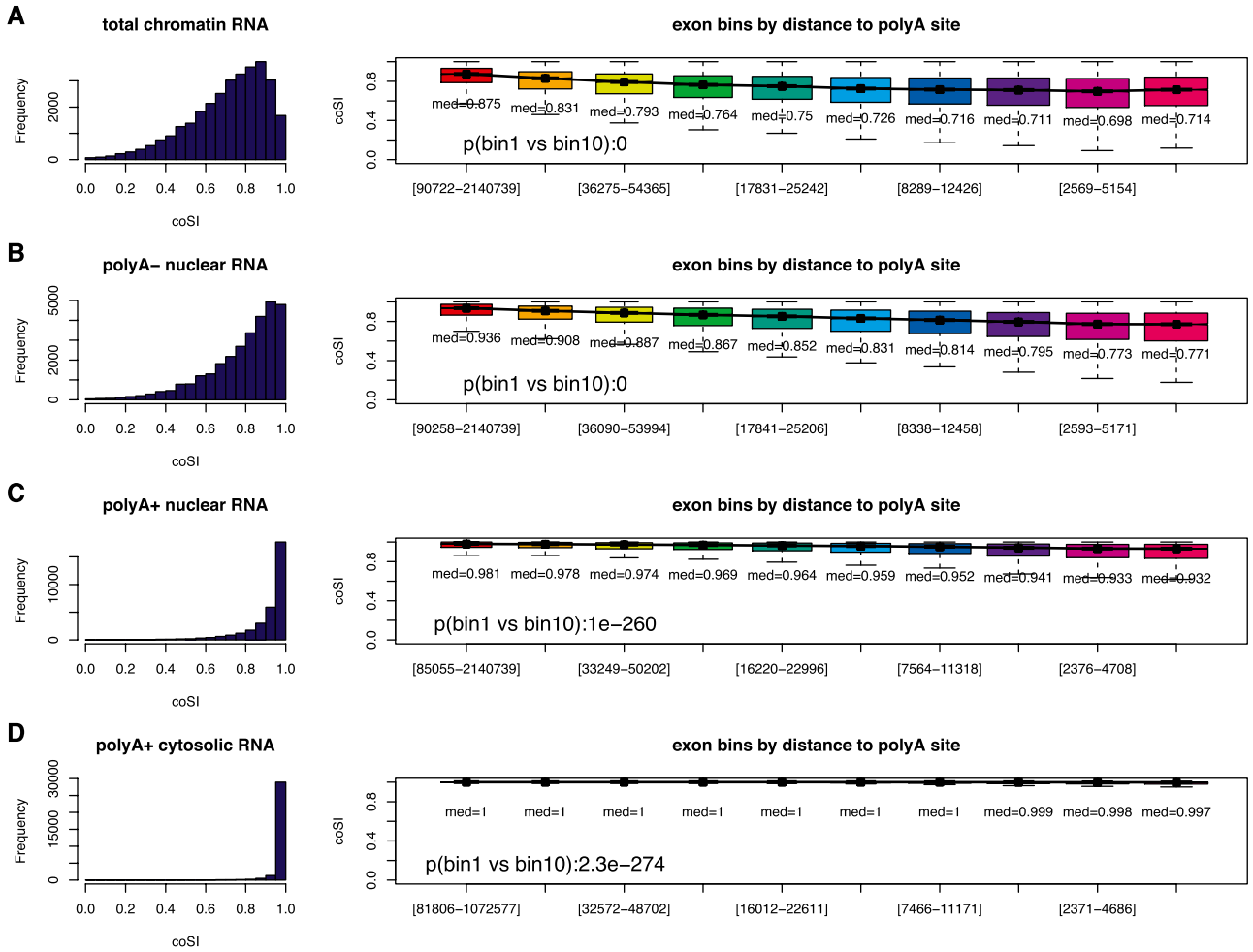

Figure 2. Histogram of coSI values (*left*) and boxplots of coSI values in bins according to the distance of an exon to the annotated polyA site (intervals on *x*-axis give minimum and maximum distance in each bin; *right*) for the total chromatin-associated RNA fraction (*A*), the polyA− nuclear fraction (*B*), the polyA+ nuclear fraction (*C*), and the polyA+ cytosolic fraction (*D*). *P*-values were calculated comparing the first and the last bin, using a two-sided Wilcoxon rank sum test. Numbers *below* boxplots indicate the median value of the according distribution.

by completely unspliced introns, further suggesting that splicing is intimately coupled and occurs almost simultaneously with transcription.

In order to exclude the possibility that large numbers of reads mapping to junctions are in fact artifactual, we have shifted the annotation by 30 bp against the transcription direction, so that in this "fake annotation," no junction is real, although practically all exons still lie within genes. As a result, the number of reads mapping to junctions has dropped 29-fold (from 11.4 million to less than 400,000), and these fake-junction reads are highly enriched in reads with two mismatches (62% compared with 7.5% for the real annotation), suggesting that these fake-junction reads contain larger numbers of false mappings. We have therefore compared chromatin-fraction coSI values for the real and for the fake annotation, using only reads (whether mapping to junctions or genome) without mismatches. The median coSI for the real annotation is 0.71, while the median coSI for the fake annotation is 0.00 (Supplemental Fig. S1H), proving that our results are not flawed by false junction mappings.

Consistent with a general strong coupling of splicing and transcription, we have found that coSI values decrease with decreasing distance to the polyA site, pointing to a "first transcribed, first spliced" trend (Fig. 2). The trend is very strong for the chromatin-associated and polyA− nuclear RNAs and is weak or absent for the polyA+ RNAs, and we have made similar observations when using an acceptor-based-definition of completed splicing (Supplemental Fig. S2). We have observed the same trend when combining all genes together and normalizing coSI scores of exons to their relative position within the gene (Supplemental Fig. S3). In these idealized genes, the coSI reaches its maximum between 20% and 30% of the gene length, indicating that splicing of introns very near the 5′ end of the gene could require a little more time. Yet from ~20%–30% of the transcript, the coSI decreases gradually. This decrease is less strong than in Figure 2, supposedly because long and short genes are considered equally, thereby minimizing the influence of exons that are very far from the polyA site.

We have specifically examined the coSI scores of exons from two genes that constitute well-studied examples of co-transcriptional splicing—the fibronectin (Cramer et al. 1997; Kadener et al. 2001; Nogues et al. 2002; de la Mata et al. 2003; Pandya-Jones and Black 2009) and *SRC* (Pandya-Jones and Black 2009) genes, and we have found that their exons have higher coSI values in the chromatin and nuclear polyA+ fractions compared with exons of other genes (Supplemental Fig. S4). This finding is not a pure consequence of gene length, as exons of 4000 longer genes show lower coSI values than fibronectin and *SRC* exons (Supplemental Fig. S4). The above observations (see Fig. 2; Supplemental Fig. S4) are not caused by incomplete annotation. Indeed, when split-mapping unmapped reads from the total chromatin fraction and excluding all exons that are within 250 bp of a potential novel splice site, we observe essentially the same trend (Supplemental Fig. S5).

Analysis of reads mapping to the genome (including exonic reads and reads deep within the introns, both of which were not used for the coSI calculation) confirms that high coSI values correspond to exons whose surrounding introns are mostly removed. Indeed, exons with low coSI in chromatin RNA show almost flat RNA-seq profiles in this RNA fraction, whereas high coSI exons show strong RNA-seq peaks on exons (Supplemental Methods; Fig. 3A,B). Exons with very low coSI values in the chromatin fraction seem to correspond to exons whose surrounding introns are spliced later, even after polyadenylation, whereas introns surrounding exons with medium or high coSI values in the total chromatin fraction seem to be spliced early, as the former show intronic reads, whereas the latter do not, in the polyA+ nuclear fraction (Fig. 3C). As expected, all exon groups show almost no evidence for intronic unspliced reads in the cytosolic polyA+ fraction (Fig. 3D). When these profiles are normalized for cytosolic polyA+ gene expression, the peak height of all three exon bins is essentially identical in the cytosolic polyA+ fraction (Supplemental Fig. S6), indicating that this observed peak height is characteristic for completed splicing. Consistent with the 5′-to-3′ coSI bias, we have found higher intronic compared with exonic read-depth as exons are closer to the polyA site (Supplemental Fig. S7).

To further rule out the possibility that our observations may originate from technical artifacts, we have analyzed CAGE tags and antisense reads, in addition to clustering the subnuclear fractions according to coSI scores. The results strongly argue against our observations originating from technical artifacts (Supplemental Information; Supplemental Figs. S8, S9).

Gene coSI values

In order to complement the exon-based view of splicing completion, we have computed coSI values at the gene level, as a function of the number of reads mapping to intron–exon junctions within the gene and of the number of reads split mapping between exons from the gene (Supplemental Information and Methods). We have found the median gene coSI value in the total chromatin fraction to be 0.618, again supporting the idea that a majority of splicing events is carried out co-transcriptionally.

Spliceosomal RNAs are enriched in chromatin-associated RNA

If splicing occurs mostly co-transcriptionally and therefore in proximity to the chromatin template, one would expect that RNAs of the splicing machinery would also reside in proximity to chromatin. We have therefore investigated the subcellular location of U1-U6 and U6atac (UxRNAs) based on RNA-seq of small RNAs performed in five different subcellular locations (Nucleus, Cytosol, Nucleoplasm, Nucleoli, and Chromatin) (Djebali et al. 2012; Supplemental Methods). As predicted, all spliceosomal UxRNAs—that is, U1, U2, U4, U5, U6, and U6atac, but not U3—are clearly enriched in the chromatin-associated fraction compared with the other fractions (Fig. 4A,B,D–G). In contrast, U3 and snoRNAs (excluding U RNAs), both of which are thought not to be involved in splicing, were highly enriched in the nucleoli fraction (Fig. 4C,H), as expected from their known functions. Of special interest in this respect is the observation that U6atac, a spliceosomal RNA of the minor spliceosome, is also enriched in the chromatin fraction. This strongly suggests two things: First, the minor spliceosome, similar to the major spliceosome, is assembled co-transcriptionally; and, second, if co-transcriptional spliceosome assembly is an attribute of both spliceosome systems, it appears likely that both types of intron removal occur in the same way, namely, co-transcriptionally in most cases.

Exon coSI values correlate with features of gene, exon, and chromatin structure

A number of sequence features characterizing the exons and their surrounding regions seem to weakly correlate with exon coSI values in chromatin (Supplemental Fig. S10). The most notable correlation is with distance to the PolyA site (see also Fig. 2) and, albeit

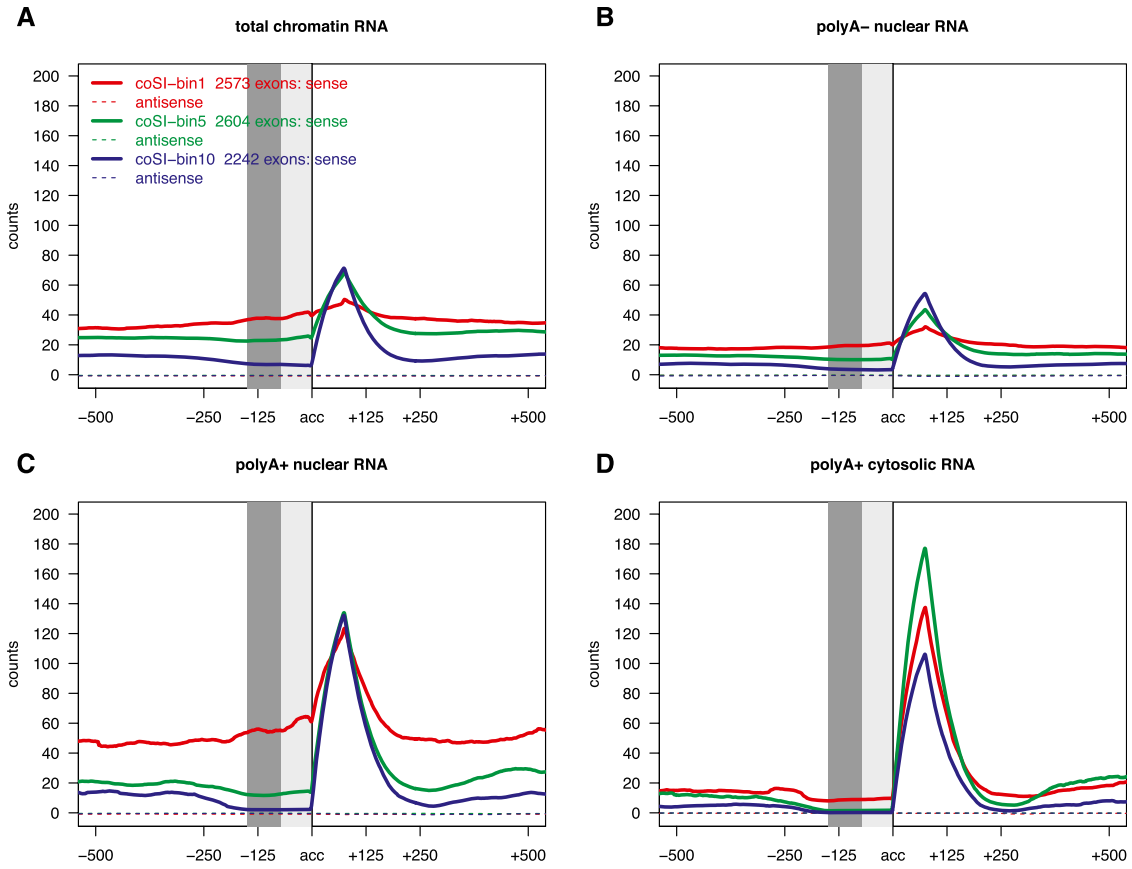

Figure 3. RNA-seq profile plots using RNA-seq reads mapping to the genome only, aligned at the acceptor. At each aligned position, the average number of overlapping RNA-seq reads (mapping to the genome) for all exons in each bin (according to coSI values in the total chromatin-associated RNA fraction in all four subfigures) is plotted for sense (solid lines) and antisense strand (dashed lines). Here, only exons that are at least 150 bp away from any other exon are used. RNA-seq profiles for the total chromatin-associated RNA fraction (*A*), the polyA− nuclear fraction (*B*), the polyA+ nuclear fraction (*C*), and the polyA+ cytosolic fraction (*D*). (Dark gray area) Positions that are guaranteed to be covered only by reads that were not used for the coSI value calculation; (both gray areas) positions that are guaranteed to be intronic. Note that these profiles are not normalized for gene expression. We added profiles normalized for cytosolic polyA+ gene expression in Supplemental Figure S6.

somehow weaker, with the distance to the transcription start site (TSS). In addition, exon coSI values correlate positively with the strength of the acceptor sites and GC content and anti-correlate with the length of the downstream intron—this is supposedly because reads spanning the exon–intron border can be observed once the donor is transcribed, while splicing can only be carried out once the entire downstream intron is transcribed. It also appears that the presence of binding sites for some splicing factors weakly correlate with coSI scores (data not shown). We have further investigated the exonic behavior of a number of chromatin modifications (Ernst et al. 2011, monitored through ChIP-seq in K562, see Supplemental Information and Methods) depending on the exon coSI value in chromatin-associated RNA. All chromatin marks monitored, as well as nucleosome (Kundaje et al. 2012) and Pol II occupancy, negatively correlate with chromatin coSI values (Supplemental Fig. S10). That is, there is a general enrichment of chromatin marks in exons with low coSI values, consistent with the DNA in these exons being still in chromatin status before or during transcription.

In order to understand how all these factors may contribute to co-transcriptional splicing, we have built a linear model in which exon coSI values in the chromatin are predicted from these factors. The linear model using all 84 variables (Supplemental Methods) achieves a correlation coefficient (cc) of 0.48 (Fig. 5A), comparing observed and predicted coSI values. The distance of the exons from annotated TSSs and polyA-sites are the most informative variables (cc of ~0.31) (Fig 5B). Acceptor and donor strength, as well as length of the surrounding introns and the exon itself, GC content of the exon, gene expression in the nuclear polyA+ fraction, chromatin status and marks, and binding sites for splicing factors progressively add more information to the prediction of coSI values (Fig. 5B–E).

coSI values reflect contrasting patterns of splicing dynamics

Analysis of coSI values across fractions reveals the specific processing pattern of the RNA in the vicinity of exons (Fig. 6A). Indeed, for 94% of the exons, the coSI value shows a monotonically increasing behavior from the total chromatin through the polyA+ nuclear to the polyA+ cytosolic fraction. Clustering according to coSI values of exons of the RNA fractions indicates that there is a large population of exons that are rapidly spliced in the chromatin fraction, while, at the other end, there is a smaller population of exons that appear to delay completion of splicing even after polyadenylation, just before exporting to the cytosol (Fig. 6A). We have specifically investigated the characteristic traits of exons that appear to delay splicing post-transcriptionally. Thus, we have arbitrarily selected

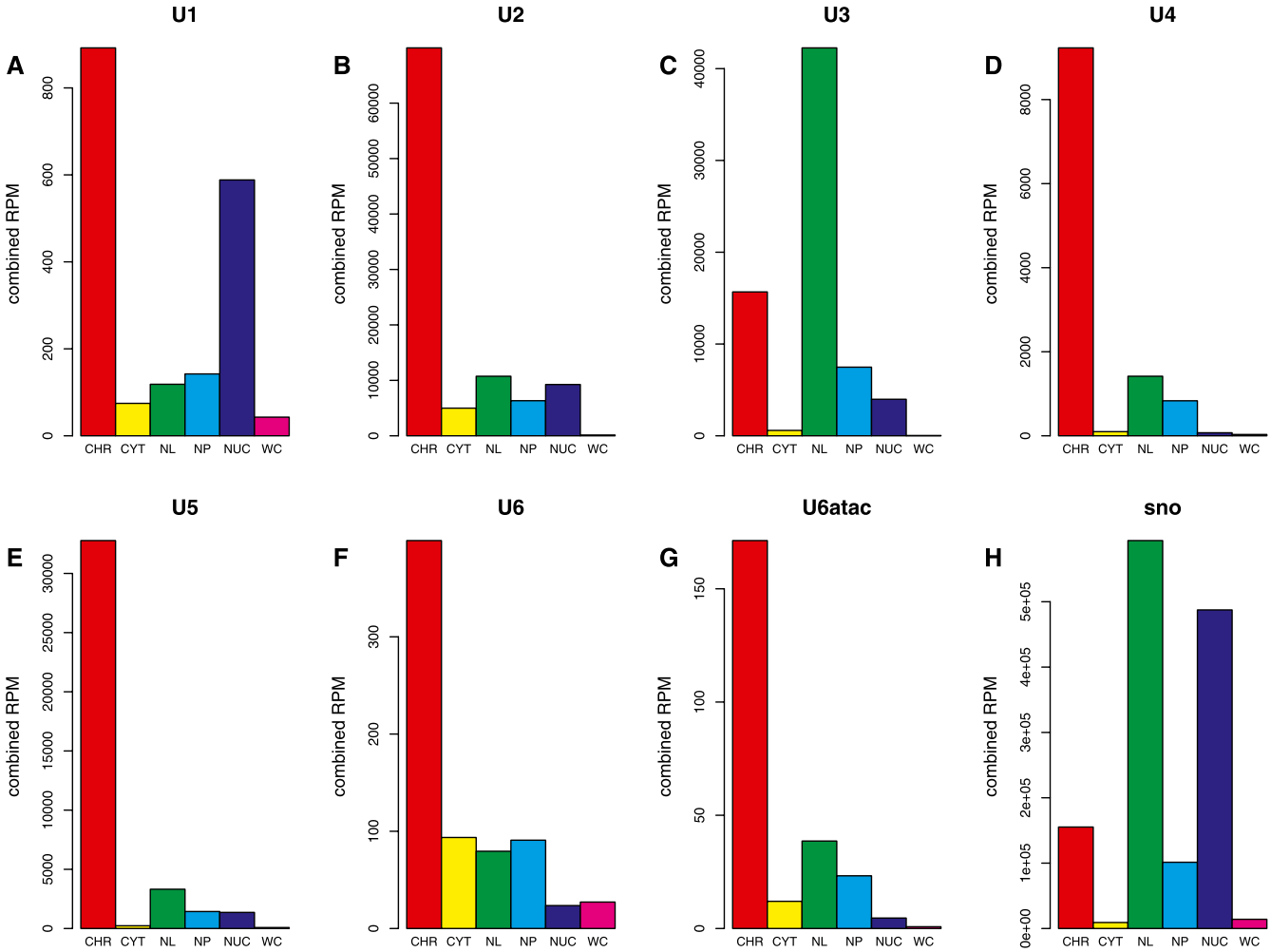

Figure 4. An RPM was calculated based on short RNA-seq in each subcellular fraction—total chromatin fraction (CHR; red), total cytoplasmic fraction (CYT; yellow), total nucleoli fraction (NL; green), total nucleoplasmic fraction (NP; light blue), total nuclear fraction (NUC; purple), total whole-cell fraction (WC; pink)—and summed for all genes encoding for U1-RNA (*A*), U2-RNA (*B*), U3-RNA (*C*), U4-RNA (*D*), U5-RNA (*E*), U6 RNA (*F*), U6atac (*G*), and non–U-RNA snoRNAs (*H*).

1390 exon candidates "with a tendency for postTS (postTS-exons)," as those with low coSI values (≤0.75) in the polyA+ nuclear fraction but high coSI values in the cytosolic polyA+ fraction (≥0.95) (Supplemental Methods). We have found that this set of postTS-exons contains 1.5-fold more alternative exons than expected by chance ($P < 3.5 \times 10^{-7}$) (Fig. 6C; see Supplemental Information and Supplemental Figures S13 and S14 for details on how alternatively spliced exons were selected). The set of postTS-exons, on the other hand, is slightly but significantly depleted of protein coding exons ($P < 5.6 \times 10^{-15}$) (Fig. 6B) and, consequently, is enriched in UTR exons—in particular in 5′UTR exons (41 out of 480 exons that are entirely within the 5′UTR, two-sided fisher P: 1.3×10^{-4}). Since this observation is in apparent contradiction with the general trend of lower coSI values toward the 3′end of the gene and also because the maximum coSI was not reached at the very beginning of the gene (Fig. S3), we have specifically investigated coSI values as a function of exon order. We have found that decreasing coSI values with increased distance (and exon order) from the TSS is only valid from the third exon on (Supplemental Fig. S11), with the second exon having slightly lower coSI values than the third exon, suggesting that the first intron is removed more slowly. This lower coSI value is paralleled by a trend for very long first introns and gradually decreasing intron size along the gene (Supplemental Fig. S12). Consistent with this interpretation, it is also known that acceptor and donor strength increase with distance to the TSS (Spies et al. 2009). An influence of the 5′ CAP (O'Mullane and Eperon 1998) could also contribute to first introns being spliced differently from other introns. All of these features would result in slower intron removal of the first intron and in lower coSI values of the second exon (Supplemental Fig. S11). On the 3′end of the gene, on the other hand, it appears that co-transcriptional intron removal, although disfavored by proximity to the polyA-site, is favored by shorter introns and stronger splice-sites, so that postTS remains relatively rare.

The observation that splicing dynamics differ between protein coding and noncoding exons has prompted us to specifically investigate the splicing dynamics of lncRNAs (see Supplemental Methods; Derrien et al. 2012). We find that coSI values of lncRNA exons, as a class, but also those of well-investigated lncRNAs (*H19*, *XIST*, *U50HG_SNHG5*) are dramatically lower than those of coding exons in the total chromatin fraction (Fig. 6D). Also in terms of completed splicing of the entire RNA, that is, on the gene-coSI

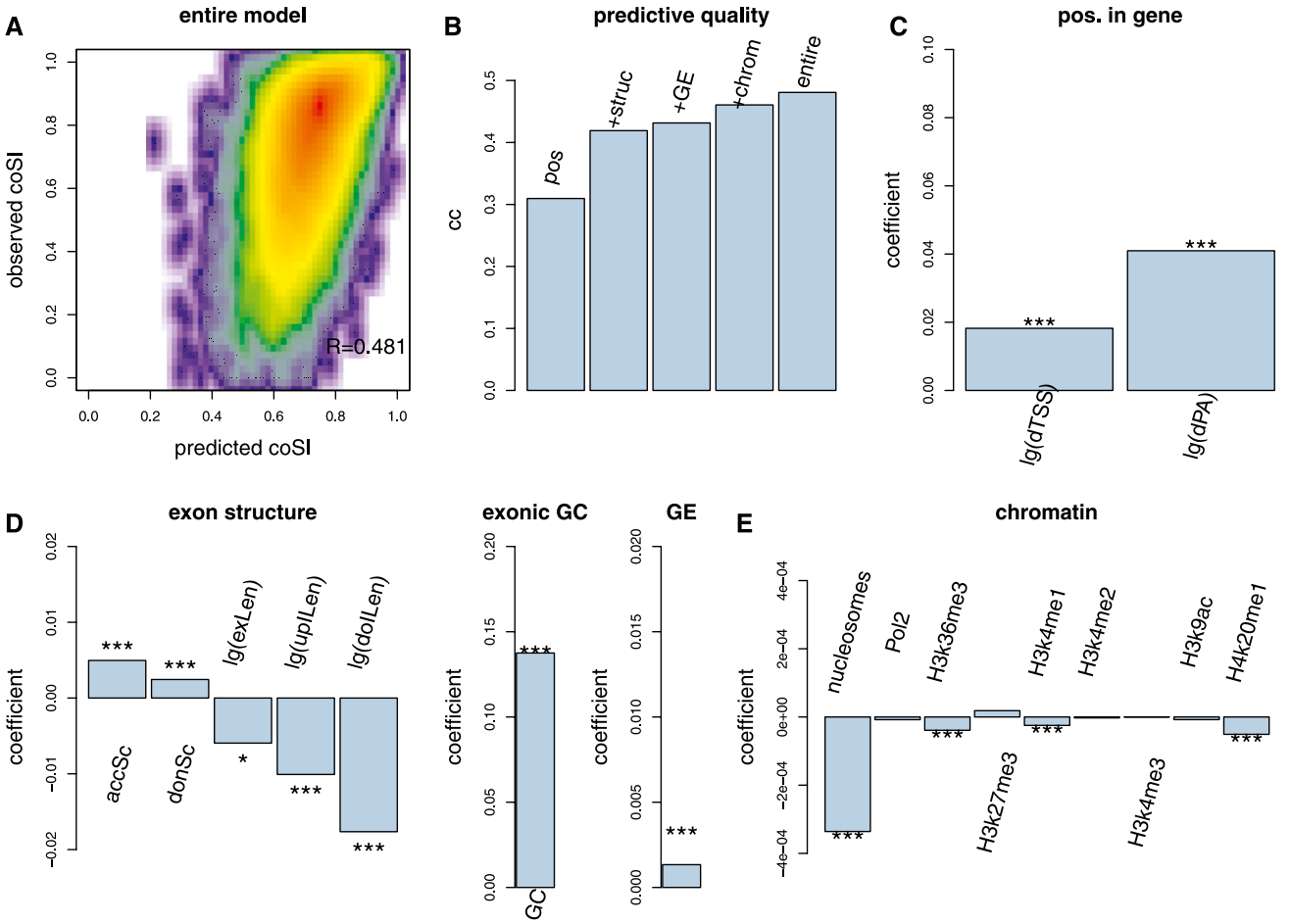

Figure 5. Linear model connecting exon-coSI values to gene, exon, and chromatin structure variables. (*A*) Smoothed scatterplot and correlation between predicted coSI values and measured coSI values using the entire model. (*B*) Correlation of predicted coSI values and measured coSI values using four increasing subsets of variables and the entire model: model with distance to TSS and distance to polyA site (pos); model additionally including acceptor strength, donor strength, log-exon-length, log-upstream-intron-length, log-downstream-intron-length and exonic GC content (+struc); model additionally including gene RPKMs from polyA+ nuclear RNA (+GE); model additionally including ChIP-seq related variables (+chrom); model including all variables (entire). (*C*) Coefficients in the entire model of distance to the TSS and to the polyA-site. (*D*) Acceptor strength (accSc), donor strength (donSc), exonic GC content (GC), log-exon-length [lg(exLen)], log-upstream-intron-length [lg(upILen)], log-downstream-intron length [lg(doIlen)] and gene RPKMs from polyA+ nuclear RNA (GE). (*E*) MNase and histone modification values as described in Figure S10.

level, lncRNAs show lower splicing completion then mRNAs in the total chromatin fraction (Fig. 6E). The difference between lncRNA exons and mRNA exons persists in the nuclear polyA+ fraction (Fig. 6F), arguing that lncRNAs are often spliced later and sometimes might even not be spliced at all. This is consistent with reports that some lncRNAs remain predominantly unspliced, for example, *AIRN* and *KCNQ1OT1* (Sleutels et al. 2002; Mancini-Dinardo et al. 2006).

Discussion

Co-transcriptional splicing has recently been shown to be widespread in the intron-poor genome of *S. cerevisiae* (Carrillo Oesterreich et al. 2010). In higher eukaryotes, co-transcriptional splicing has been documented in detail for a few individual genes such as fibronectin and *SRC* (Cramer et al. 1997; Kadener et al. 2001; de la Mata et al. 2003; Pandya-Jones and Black 2009), and this mode of intron removal has been proposed to be widespread in the human brain, based on analysis of whole-cell, total RNA-seq (Ameur et al. 2011). While we coincide on the claim of widespread co-transcriptional splicing, our approach of analyzing RNA-seq data in a variety of fractions provides major advantages: First, we are able to clearly separate cytosolic RNAs, nuclear RNAs, as well as a special subset of the latter, RNAs that still reside on the chromatin template. Thus we demonstrate that spliced reads and exonic reads in the latter fraction are not the result of completely spliced RNAs, which still remain in cytosol (or nucleus) and not on the chromatin. Second, we can define, for single exons, what proportion of their surrounding introns is removed co-transcriptionally or after polyadenylation while controlling for intron retention with cytosolic RNA-seq. This we can achieve, by introducing an exon-based measure of splicing completion: This measure, the coSI, shows that most introns initiate splicing while the RNA is still associated with the chromatin—strongly suggesting that co-transcriptional splicing is also the dominant mode in the human genome. Consistent with this, we have found significant enrichment of spliceosomal snRNAs in chromatin-associated RNA compared with other cellular RNA fractions and other nonspliceosomal snRNAs. This supports the idea that exons, around which we detect a tendency for postTS, might already have been committed to splicing co-transcriptionally. Al-

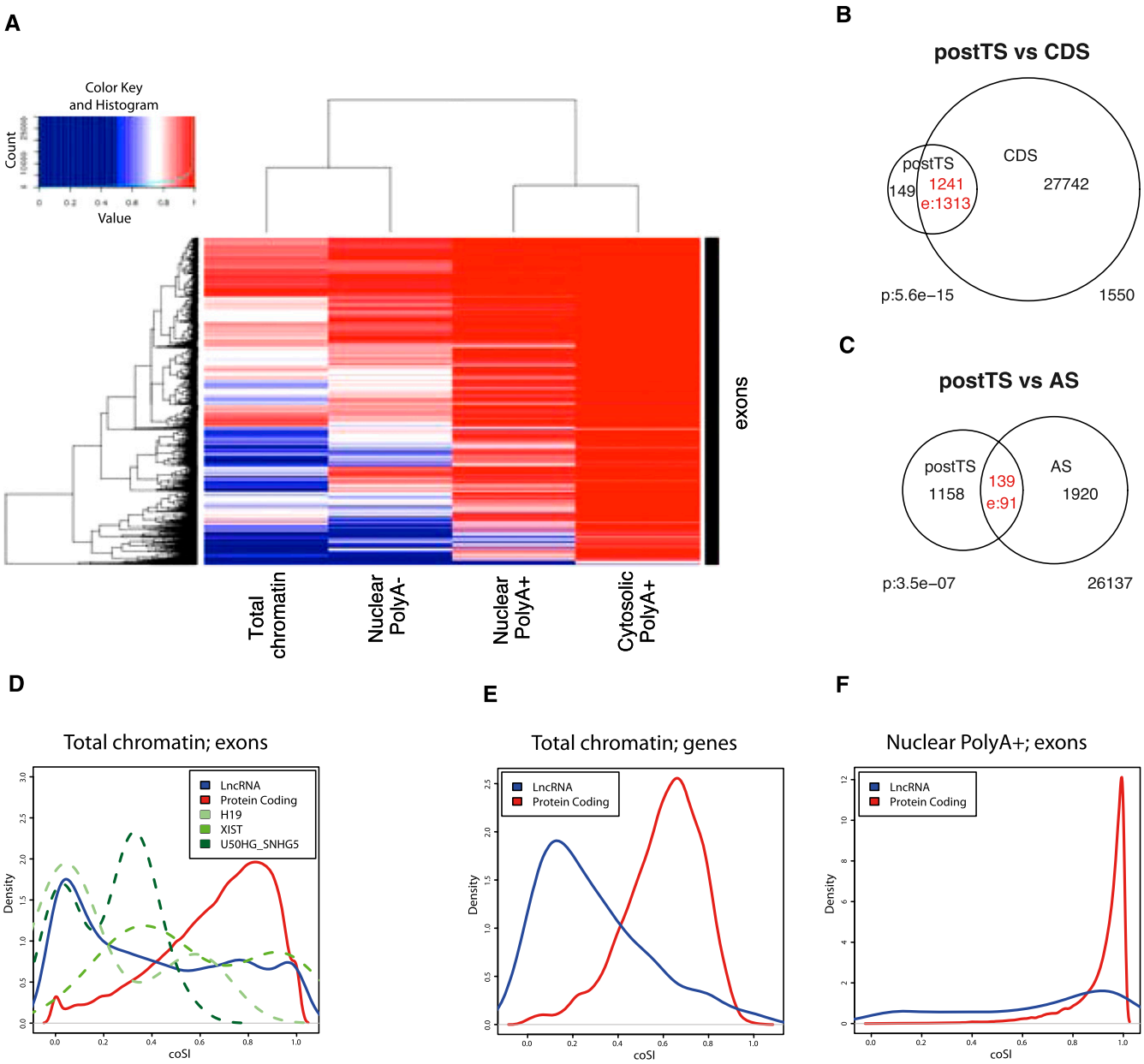

Figure 6. (*A*) Clustering of subcellular RNA fractions and exons according to exonic coSI values using four RNA fractions. From *left* to *right*: total chromatin-associated RNA, polyA− nuclear RNA, polyA+ nuclear RNA, polyA+ cytosolic RNA. Note that the scale is only linear from coSI ≥ 0.5 on. (*B*) Overlap between exons with a tendency for post-transcriptional splicing (postTS) and entirely coding exons (CDS). (*C*) Overlap between cell type specifically included AS-exons and exons with a tendency for postTS. (*D*) The distribution of coSI scores for various exon sets is shown, based on calculations for chromatin total RNA. Information is plotted for the 4933 lncRNA exons and 372,306 protein-coding gene exons that have sufficient RNA-seq reads to calculate a confident coSI score. In addition, we extracted exon values for three known lncRNAs: *H19* (18 exons), *XIST* (19 exons), *U50HG-SNHG5* (22 exons). The difference between lncRNA and protein exon coSI values is statistically significant (Wilcox test; $P < 2.2 \times 10^{-16}$). (*E*) Gene-level coSI scores from chromatin total RNA are plotted for 92 lncRNAs and 4066 protein-coding genes. The difference between the distributions is statistically significant (Wilcox test; $P < 2.2 \times 10^{-16}$). (*F*) Exon-level coSI scores from nuclear polyA+ RNA are plotted for 206 lncRNA exons and 32,496 protein-coding exons. The difference between the distributions is statistically significant (Wilcox test; $P < 2 \times 10^{-16}$).

though strictly speaking we cannot detect this commitment on the premessenger RNA, the contrasting behavior of snoRNAs and U3 snRNAs, compared with other spliceosomal snRNAs, is highly suggestive. Indeed, it has been shown that the elongation rate can affect inclusion of exon E33 of the fibronectin gene without affecting the relative order in which introns are removed (de la Mata et al. 2010). A corollary of this observation is that, in this case, commitment to inclusion can be achieved co-transcriptionally, while actual intron removal might occur later (de la Mata et al. 2010). Hence transcription-mediated influences on splicing are probably larger than can be detected with the data analyzed here.

A variety of recent studies have linked chromatin structure to splicing. Co-transcriptionality of splicing is not an absolute prerequisite for a chromatin-splicing connection, because chromatin

could influence commitment rather than actual intron removal (see above). In the light of our data, it seems, however, that the majority of splicing occurs during transcription and thereby offers an even more direct opportunity for chromatin to influence splicing. Indeed, we detect enrichment of a variety of chromatin marks on exons in the process of being spliced (i.e., exons, with low coSI values in the chromatin-associated RNA).

While proximity to the polyA site seems to disfavor co-transcriptional splicing near the end of the gene, other features such as 5′-to-3′ decreasing intron size and increasing splice site strength favor rapid splicing, so that comparatively high co-transcriptional splicing completion can still be observed toward the 3′ end of the gene. Moreover, it is known that various histone marks vary along the gene (Barski et al. 2007). Such a special chromatin organization toward the 3′ end of the gene could also contribute either directly or indirectly to splicing completion prior to polyadenylation.

Interestingly, gene expression of nuclear polyA+ RNAs is a weak but significant predictor of coSI values in the chromatin total fraction, suggesting a selective pressure for splicing in more highly expressed genes occurs more rapidly. Splicing around coding exons is significantly more often co-transcriptional (in comparison to exons containing noncoding sequence), while splicing around alternatively skipped exons is significantly more post-transcriptional (than for exons not involved in skipping events). Importantly, this does not imply that for all alternative exons splicing of the corresponding introns always occurs post-transcriptionally. Rather, it means that while only a few introns surrounding constitutive exons are removed post-transcriptionally, a significantly higher fraction of introns surrounding (or skipping) alternative exons are removed post-transcriptionally. This genome-wide picture supposedly represents a mixture of two models observed on the fibronectin gene (de la Mata et al. 2010) and on the *Sxl* and *PTBP2* (also known as *nPTB*) genes (Vargas et al. 2011). In the former case, changed exon inclusion levels were achieved without changing the relative timing of actual intron removal but, supposedly, rather by changing splicing commitment co-transcriptionally (de la Mata et al. 2010). In the latter, however, exon inclusion occurred when splicing was co-transcriptional, whereas the exon was skipped when splicing was carried out post-transcriptionally (Vargas et al. 2011). One interpretation for these and our observations is that co-transcriptional splicing tends to be more faithful than postTS, which would therefore offer more opportunities for an exon to be alternatively, that is, differently, included. An interesting corollary of this idea is that when a shorter-than-usual isoform of a gene is expressed, some introns around internal exons might be spliced more often post-transcriptionally, as they are closer to the chosen polyA site. This could then lead to changed inclusion rates of the exon.

Lower coSI values for lncRNAs can be interpreted in multiple ways, all of which probably apply to different subsets of this rather heterogeneous RNA class. Some splicing events in lncRNAs are probably carried out later, that is, post-transcriptionally, simply because lncRNA gene features (e.g., shorter gene length, lower expression) favor this mode of splicing. It is highly likely given the data presented here and previously described examples, that many lncRNAs either (1) remain completely unspliced or (2) have a high proportion of primary transcripts that are never spliced, while a minority are processed by the splicing machinery. For example, two lncRNAs involved in imprinting are likely to remain in the nucleus in an unspliced state: *AIRN* (Sleutels et al. 2002) and *KCNQ1OT1* (Mancini-Dinardo et al. 2006). However, our data should be treated with caution, since the analysis was carried out on the small subset of lncRNAs that are expressed sufficiently highly to calculate a coSI score with confidence (see Supplemental Methods). Nevertheless, the coSI data presented here will be a valuable tool for subclassifying lncRNAs by their processing status.

In summary, we believe that our results strongly suggest that splicing is a highly co-transcriptional process, whose outcome depends crucially on many factors in the exon, and the overall gene sequence, as well as on chromatin architecture and transcription dynamics. As our analysis reveals here, the interrogation of RNA fractions provides invaluable information on the processing pathways establishing RNA genealogy.

Data access

Supplemental Table S1 can be accessed at http://genome.crg.es/~htilgner/2011_coSI_paper/2011cp_index.html. Raw RNA-seq reads can be accessed at the NCBI Gene Expression Omnibus (GEO) (http://www.ncbi.nlm.nih.gov/geo/) under accession numbers GSE30567 and GSE24565. Additional detailed methods for RNA-seq can be obtained in the production documents under "CSHL Long RNA-seq" and "CSHL Sm RNA-seq" at http://genome.ucsc.edu/ENCODE/downloads.html.

Competing interest statement

Michael Snyder is a consultant for Illumina and on the scientific advisory board of Personalis and GenapSys.

Acknowledgments

We thank Juan Valcárcel and Tobias Warnecke from the CRG for useful discussions. This work has been carried out under grants RD07/0067/0012, BIO2006-03380, and CSD2007-00050 from the Spanish Ministry of Science, and grants 1U54HG004557-01 and 1U54HG004555-01 from the National Institutes of Health.

References

Allo M, Buggiano V, Fededa JP, Petrillo E, Schor I, de la Mata M, Agirre E, Plass M, Eyras E, Elela SA, et al. 2009. Control of alternative splicing through siRNA-mediated transcriptional gene silencing. *Nat Struct Mol Biol* **16:** 717–724.

Ameur A, Zaghlool A, Halvardson J, Wetterbom A, Gyllensten U, Cavelier L, Feuk L. 2011. Total RNA sequencing reveals nascent transcription and widespread co-transcriptional splicing in the human brain. *Nat Struct Mol Biol* **18:** 1435–1440.

Andersson R, Enroth S, Rada-Iglesias A, Wadelius C, Komorowski J. 2009. Nucleosomes are well positioned in exons and carry characteristic histone modifications. *Genome Res* **19:** 1732–1741.

Barash YCJ, Gao W, Pan Qu, Wang X, Shai O, Blencowe J, Frey B. 2010. Deciphering the splicing code. *Nature* **465:** 53–59.

Barski A, Cuddapah S, Cui K, Roh TY, Schones DE, Wang Z, Wei G, Chepelev I, Zhao K. 2007. High-resolution profiling of histone methylations in the human genome. *Cell* **129:** 823–837.

Beyer AL, Osheim YN. 1988. Splice site selection, rate of splicing, and alternative splicing on nascent transcripts. *Genes Dev* **2:** 754–765.

Carrillo Oesterreich F, Preibisch S, Neugebauer KM. 2010. Global analysis of nascent RNA reveals transcriptional pausing in terminal exons. *Mol Cell* **40:** 571–581.

Cramer P, Pesce CG, Baralle FE, Kornblihtt AR. 1997. Functional association between promoter structure and transcript alternative splicing. *Proc Natl Acad Sci* **94:** 11456–11460.

de la Mata M, Alonso CR, Kadener S, Fededa JP, Blaustein M, Pelisch F, Cramer P, Bentley D, Kornblihtt AR. 2003. A slow RNA polymerase II affects alternative splicing in vivo. *Mol Cell* **12:** 525–532.

de la Mata M, Lafaille C, Kornblihtt AR. 2010. 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.

Derrien T, Johnson R, Bussotti G, Tanzer A, Djebali S, Tilgner H, Guernec G, Martin D, Merkel A, Knowles DG, et al. 2012. The GENCODE v7 catalog of human long noncoding RNAs: Analysis of their gene structure, evolution, and expression. *Genome Res* (this issue). doi: 10.1101/gr.132159.111.

Djebali S, Davis CA, Merkel A, Dobin A, Lassmann T, Mortazavi AM, Tanzer A, Lagarde J, Lin W, Schlesinger F, et al. 2012. Landscape of transcription in human cells. *Nature* (in press).

Ernst J, Kheradpour P, Mikkelsen TS, Shoresh N, Ward LD, Epstein CB, Zhang X, Wang L, Issner R, Coyne M, et al. 2011. Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature* **473:** 43–49.

Graveley BR. 2000. Sorting out the complexity of SR protein functions. *RNA* **6:** 1197–1211.

Hon G, Wang W, Ren B. 2009. Discovery and annotation of functional chromatin signatures in the human genome. *PLoS Comput Biol* **5:** e1000566. doi: 10.1371/journal.pcbi.1000566.

Howe KJ, Kane CM, Ares M Jr. 2003. Perturbation of transcription elongation influences the fidelity of internal exon inclusion in *Saccharomyces cerevisiae*. *RNA* **9:** 993–1006.

Kadener S, Cramer P, Nogues G, Cazalla D, de la Mata M, Fededa JP, Werbajh SE, Srebrow A, Kornblihtt AR. 2001. Antagonistic effects of T-Ag and VP16 reveal a role for RNA pol II elongation on alternative splicing. *EMBO J* **20:** 5759–5768.

Kolasinska-Zwierz P, Down T, Latorre I, Liu T, Liu XS, Ahringer J. 2009. Differential chromatin marking of introns and expressed exons by H3K36me3. *Nat Genet* **41:** 376–381.

Kundaje A, Kyriazopoulou-Panagiotopoulou S, Libbrecht M, Smith CL, Raha D, Winters EE, Johnson SM, Snyder M, Batzoglou S, Sidow A. 2012. Ubiquitous heterogeneity and asymmetry of the chromatin environment at regulatory elements. *Genome Res* (this issue). doi: 10.1101/gr.136366.111.

Luco RF, Pan Q, Tominaga K, Blencowe BJ, Pereira-Smith OM, Misteli T. 2010. Regulation of alternative splicing by histone modifications. *Science* **327:** 996–1000.

Mancini-Dinardo D, Steele SJ, Levorse JM, Ingram RS, Tilghman SM. 2006. Elongation of the Kcnq1ot1 transcript is required for genomic imprinting of neighboring genes. *Genes Dev* **20:** 1268–1282.

Nahkuri S, Taft RJ, Mattick JS. 2009. Nucleosomes are preferentially positioned at exons in somatic and sperm cells. *Cell Cycle* **8:** 3420–3424.

Nogues G, Kadener S, Cramer P, Bentley D, Kornblihtt AR. 2002. Transcriptional activators differ in their abilities to control alternative splicing. *J Biol Chem* **277:** 43110–43114.

O'Mullane L, Eperon IC. 1998. The pre-mRNA 5′ cap determines whether U6 small nuclear RNA succeeds U1 small nuclear ribonucleoprotein particle at 5′ splice sites. *Mol Cell Biol* **18:** 7510–7520.

Pandya-Jones A, Black DL. 2009. Co-transcriptional splicing of constitutive and alternative exons. *RNA* **15:** 1896–1908.

Roberts GC, Gooding C, Mak HY, Proudfoot NJ, Smith CW. 1998. Co-transcriptional commitment to alternative splice site selection. *Nucleic Acids Res* **26:** 5568–5572.

Schor IE, Rascovan N, Pelisch F, Allo M, Kornblihtt AR. 2009. Neuronal cell depolarization induces intragenic chromatin modifications affecting NCAM alternative splicing. *Proc Natl Acad Sci* **106:** 4325–4330.

Schwartz S, Meshorer E, Ast G. 2009. Chromatin organization marks exon-intron structure. *Nat Struct Mol Biol* **16:** 990–995.

Shukla S, Kavak E, Gregory M, Imashimizu M, Shutinoski B, Kashlev M, Oberdoerffer P, Sandberg R, Oberdoerffer S. 2011. CTCF-promoted RNA polymerase II pausing links DNA methylation to splicing. *Nature* **479:** 74–79.

Sleutels F, Zwart R, Barlow DP. 2002. The non-coding Air RNA is required for silencing autosomal imprinted genes. *Nature* **415:** 810–813.

Smith CW, Valcárcel J. 2000. Alternative pre-mRNA splicing: The logic of combinatorial control. *Trends Biochem Sci* **25:** 381–388.

Spies N, Nielsen CB, Padgett RA, Burge CB. 2009. Biased chromatin signatures around polyadenylation sites and exons. *Mol Cell* **36:** 245–254.

Tilgner H, Nikolaou C, Althammer S, Sammeth M, Beato M, Valcarcel J, Guigo R. 2009. Nucleosome positioning as a determinant of exon recognition. *Nat Struct Mol Biol* **16:** 996–1001.

Vargas DY, Shah K, Batish M, Levandoski M, Sinha S, Marras SA, Schedl P, Tyagi S. 2011. Single-molecule imaging of transcriptionally coupled and uncoupled splicing. *Cell* **147:** 1054–1065.

Wang Z, Burge CB. 2008. Splicing regulation: From a parts list of regulatory elements to an integrated splicing code. *RNA* **14:** 802–813.

Wang Z, Rolish ME, Yeo G, Tung V, Mawson M, Burge CB. 2004. Systematic identification and analysis of exonic splicing silencers. *Cell* **119:** 831–845.

Received November 6, 2011; accepted in revised form February 7, 2012.