Connections between Alternative Transcription and Alternative Splicing in Mammals

Connections between Alternative Transcription and Alternative Splicing in Mammals

Svetlana A. Shabalina*,[1], Alexey N. Spiridonov[2], Nikolay A. Spiridonov[3], and Eugene V. Koonin*,[1]

[1]National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, Maryland

[2]Department of Mathematics, Massachusetts Institute of Technology

[3]Division of Therapeutic Proteins, Center for Drug Evaluation and Research, US Food and Drug Administration, Bethesda, Maryland

*Corresponding author: E-mail: shabalin@ncbi.nlm.nih.gov; koonin@ncbi.nlm.nih.gov.

Accepted: 27 September 2010

Abstract

The majority of mammalian genes produce multiple transcripts resulting from alternative splicing (AS) and/or alternative transcription initiation (ATI) and alternative transcription termination (ATT). Comparative analysis of the number of alternative nucleotides, isoforms, and introns per locus in genes with different types of alternative events suggests that ATI and ATT contribute to the diversity of human and mouse transcriptome even more than AS. There is a strong negative correlation between AS and ATI in 5′ untranslated regions (UTRs) and AS in coding sequences (CDSs) but an even stronger positive correlation between AS in CDSs and ATT in 3′ UTRs. These observations could reflect preferential regulation of distinct, large groups of genes by different mechanisms: 1) regulation at the level of transcription initiation and initiation of translation resulting from ATI and AS in 5′ UTRs and 2) posttranslational regulation by different protein isoforms. The tight linkage between AS in CDSs and ATT in 3′ UTRs suggests that variability of 3′ UTRs mediates differential translational regulation of alternative protein forms. Together, the results imply coordinate evolution of AS and alternative transcription, processes that occur concomitantly within gene expression factories.

Key words: alternative splicing, alternative transcription initiation, alternative transcription termination, gene expression factories.

Introduction

The extraordinary complexity of transcriptomes that underpins the structural and functional diversity of mammalian proteomes is created by alternative splicing (AS) and alternative transcription (Sultan et al. 2008; Wilhelm et al. 2008). Transcriptome analysis shows that the majority of protein-coding genes in mammals undergo AS whereby the same sequence belongs to an exon in one subset of transcripts of the given gene locus and to an intron in another subset of transcripts (Blencowe 2006; Kim et al. 2008). Indeed, the latest estimates using high-throughput sequencing methods indicate that up to 95% of multiexon human genes are subject to AS and reveal approximately 100,000 major AS events (Pan et al. 2008). In addition, recent studies of mammalian gene expression point to the wide spread of alternative initiation and alternative termination of transcription (ATI and ATT, respectively) and importance of these events in the generation of the transcriptome diversity (Landry et al. 2003; Shabalina and Spiridonov 2004; Baek et al. 2007; Ma et al. 2009; Yamashita et al. 2010).

The prevalence and functional significance of different types of alternative events (AEs) differs between parts (functional domains) of transcripts. Thus, AS is common in the 5′ untranslated regions (5′ UTRs) and coding sequences (CDSs), with a significantly greater fraction of nucleotides involved in AS in the 5′ UTRs compared with the CDS (Resch et al. 2004, 2009; Cenik et al. 2010). In contrast, AS is rare in 3′ UTRs given the overall low intron density in this region (Hong et al. 2006; Grillo et al. 2010). In contrast, ATI and ATT are confined, respectively, to the 5′ UTRs and 3′ UTR and the corresponding "grey areas," the sequences that alternate between the CDS and UTRs in alternative transcripts.

Numerous biochemical and cytological experiments indicate that in eukaryotes transcription and mRNA processing including capping, splicing, and polyadenylation/cleavage form a network of elaborately regulated and coupled processes that occur together within nuclear "gene expression

factories" (Bentley 2002, 2005; Maniatis and Reed 2002; Kornblihtt et al. 2004). These findings suggest intriguing possibility that AEs occurring at different levels of gene expression and transcript processing might not be independent.

Given the wide spread of AEs in mammalian genes and the increasingly apparent transcription-splicing coupling, we undertook a genome-wide survey of the relative contributions of different types of AEs to the diversity of the transcriptomes and the connections between alternative

Fig. 1—Anatomy of mammalian transcripts: functional domains, constitutive and alternative nucleotides, and AEs. TI, transcription initiation site; TS, translation initiation site; TT, transcription termination site; *, translation termination site. Protein-coding regions are filled. Frequent AEs and common combinations of AEs are shown in red. Rare AEs and avoided combinations of AEs are shown in blue.

transcribed from a gene locus. Alternative nucleotides in each transcript that belong to the first exon located between the most upstream and downstream transcription start sites were attributed to ATI. Similarly, alternative nucleotides in each transcript that belong to the last exon located between the most upstream and downstream transcription termination sites were attributed to ATT. The remaining alternative nucleotides were considered as resulting from AS. We evaluated relaxed (100–500,000 nt) and more stringent (300–50,000 nt) thresholds for differences between the positions of the upstream and downstream ATI (ATT) sites in gene loci and found that different thresholds yielded similar results (table 2, supplementary table S2*A*–*B*, Supplementary Material online). The results are presented for the 100 nt minimum and 500,000 nt maximum (table 2, supplementary table S2*B*, Supplementary Material online), as well as for the most conservative thresholds, 300 nt minimum and 50,000 nt maximum (supplementary table S2*A*, Supplementary Material online).

As an additional control for the reliability of the AE classification, we compared the lists of gene loci from UCSC and RefSeq databases that are classified as employing ATI against the database of experimentally verified transcription start sites, DBTSS (Wakaguri et al. 2008). Approximately, 70% of gene loci from RefSeq, identified in our analysis as involved in ATI, were on the list of genes with alternative transcription starts from DBTSS database (supplementary table S3*A*, Supplementary Material online). The proportion of gene loci with experimentally validated ATI was somewhat lower for UCSC (supplementary table S3*B*, Supplementary Material online) than it was in the case of RefSeq, as one would expect given that many transcripts included in the UCSC database are predictions.

Gene Ontology Annotation

Functional annotation for human and mouse was downloaded from the Gene Ontology (GO) database (Harris et al. 2004). Starting with a total of 16,468 annotated human genes, GO annotations were mapped to 89% of the genes in the UCSC subsets. With 17,480 annotated mouse genes, GO annotations were mapped to 94% of the genes in the UCSC subsets. The GO terms associated with each group of human genes employing different types of AEs were identified and analyzed using the GoMiner program and the UniProtKB protein data set, false discovery rate (FDR) cutoff of 0.05, and 100 GoMiner runs to estimate FDR (Zeeberg et al. 2003). Keyword frequencies were tabulated for all analyzed subsets and normalized by the total numbers of genes in each set (Resch et al. 2009). *P* values were calculated using the χ^2 test.

Results

Alternative and Constitutive Nucleotides in Different Functional Domains of Mammalian Transcripts

We performed a genome-wide census of AE in human and mouse transcripts available from the UCSC and RefSeq

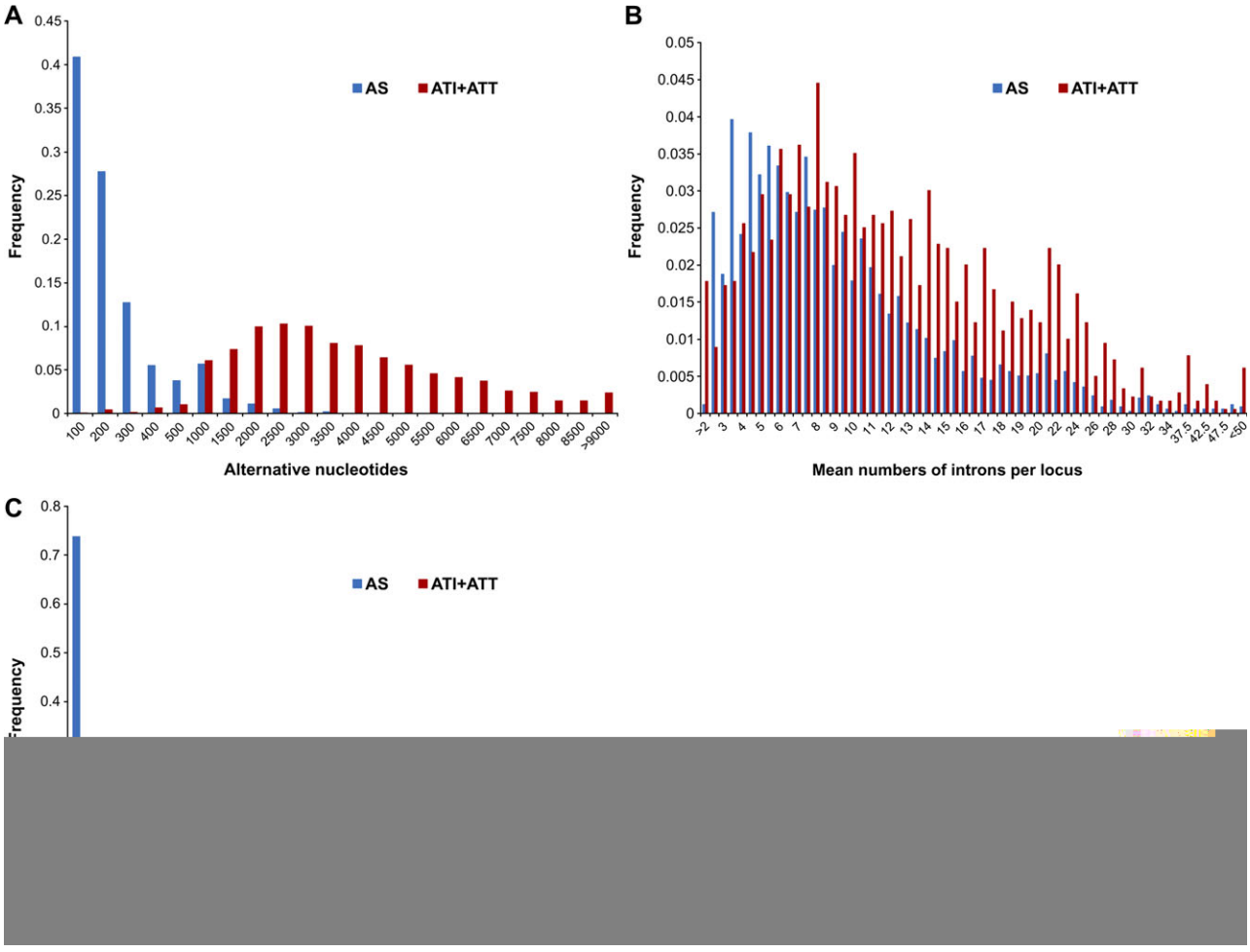

Fig. 2—Distribution of number of alternative nucleotides (*A*), introns (*B*), and isoforms (*C*) per gene locus for sets of human genes with different types of AEs. AS, gene loci with AS alone, where all transcripts from the gene locus begin and end at the same positions; ATI + ATT, gene loci with both ATI and ATT.

coupling in 5′ UTRs could be explained by the requirement of short 5′ UTRs for efficient initiation of translation (Wegrzyn et al. 2008), so that utilization of an upstream transcription start is compensated by AS.

Coupling between ATT and AS in 3′ UTRs. We also found a strong positive correlation between ATT and AS in 3′ UTRs (fig. 3, supplementary tables S6 and S7, Supplementary Material online): AS in 3′ UTRs, although rare,

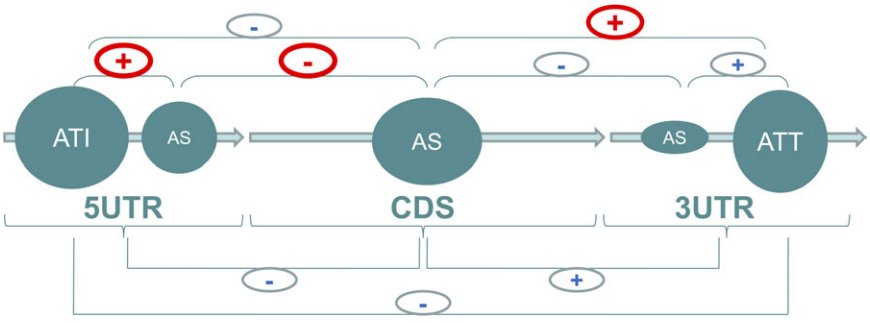

Fig. 3—.Relationships between AEs within and between the functional domains of mammalian transcripts. The diameters of the circles are roughly proportional to the prevalence of the respective AEs in the given transcript domain; (+) denotes a significant positive correlation and (−) denotes a significant negative correlation; the strongest correlations are shown in red.

was almost invariably accompanied by ATT. The expected co-occurrence frequency of ATI and AS is $0.406 \times 0.037 = 0.015$, whereas the observed co-occurrence frequency was 0.033, 2-fold greater than expected by chance and highly statistically significant ($\chi^2 = 432$, $P(\chi^2) = 2.5 \times 10^{-93}$).

Connections between Alternative Events in Different Functional Domains of Transcripts

In the preceding section, we described the apparent coupling between different types of AEs within the same domain of transcripts, such as the coupling between ATI and AS in 5′ UTRs. To gain further insight into the relationships between ATI, ATT, and AS, we examined the connections between AEs that co-occur in different functional domains, that is, in 5′ UTRs and CDSs, in CDSs and 3′ UTRs, and in both UTRs.

ATI in 5′ UTR Versus AS in CDS. There are strong, highly significant and consistent negative correlations between AEs in 5′ UTR and CDS (fig. 3 and supplementary tables S6 and S7, Supplementary Material online). Genes with AS in the CDS are substantially less likely to harbor ATI or AS in the 5′ UTR ($P < 10^{-6}$, table 2 and supplementary table S7, Supplementary Material online). The strongest negative correlation between AS in 5′ UTRs and CDS was observed for the group of genes that do not employ ATI or ATT ("AS" in table 2 and supplementary table S7, Supplementary Material online), suggestive of mutually exclusive AS in these transcript domains. This mutual avoidance of AS in CDSs and 5′ UTRs was highly significant for the complete set of transcripts from UCSC database ($\chi^2 = 1{,}205$, $P(\chi^2) = 5.7 \times 10^{-201}$, supplementary table S6, Supplementary Material online). This effect was even more striking for genes with a single transcription initiation site (without ATI). The frequency of co-occurrence of AS in 5′ UTR and CDS in this gene group was 5-fold lower than expected. The correlation between ATI in 5′ UTRs and AS in CDS was also negative and highly significant ($\chi^2 = 553$, $P(\chi^2) = 1.2 \times 10^{-119}$) but substantially weaker than the negative correlation between AS in these domains.

Coupling between ATT in 3′ UTRs and AS in CDS. We observed a consistent positive correlation between AEs in CDS, the 3′-grey areas and 3′ UTRs (fig. 3 and supplementary tables S6 and S7, Supplementary Material online). AS in the CDS is frequently accompanied with AEs in the 3′ grey area and in the 3′ UTR as well ($P < 10^{-5}$, supplementary table S7, Supplementary Material online). Moreover, ATT occurs at 3′ regions almost exclusively when there is AS in the CDS ($\chi^2 = 1390$, $P(\chi^2) = 5.5 \times 10^{-301}$; supplementary table S6, Supplementary Material online). This tight connection suggests that the variability of 3′ UTRs is functionally related to the variability of protein-CDSs.

Thus, our results reveal differential connections of AEs in 5′ UTRs and 3′ UTRs with AEs in the protein-coding regions.

AEs in Different GO Categories

A comparison of the GO categories associated with genes undergoing ATI, on the one hand, and genes undergoing AS but not ATI in the CDS, on the other hand, revealed notable, statistically significant differences. Specifically, the ATI group was enriched for genes involved in developmental processes, signal transduction, and apoptosis, whereas the AS group was enriched for genes involved in cellular processes and organization, protein modification, and regulation of metabolism (supplementary table S8, Supplementary Material online). These findings seem to support the conclusion that ATI in 5′ UTRs and AS in the CDS are differentially employed to regulate different functional classes of genes. Furthermore, it appears plausible that transcription from alternative promoters is predominantly used by tissue and/or developmental stage-specific genes, whereas AS increases diversity of protein isoforms that perform more general cellular and metabolic functions.

Discussion

It is often assumed that AS is the primary source of transcript diversity in mammals. The present analysis shows that this view is valid only for the CDS, whereas in the 5′ UTRs and the 3′ UTRs, the dominant AEs are, respectively, ATI and ATT. Our comparative analysis of alternative nucleotides, mean numbers of isoforms, and introns per locus in gene subsets with different types of AEs demonstrates that ATI and ATT contribute to the diversity of mammalian transcriptome even more than AS. We detected two types of coupling between different classes of AEs: within functional domains of transcript and between domains. In the 5′ UTRs, ATI and AS are positively correlated, revealing an unexpected dependence between two classes of a priori independent AEs. As for between domain connections, there is a tight coupling between AS in CDS and ATT in 3′ UTRs but, in contrast, a strong anticorrelation between ATI and especially AS in 5′ UTRs and AS in the CDS (fig. 3). Recent studies on connections between AS and AT in mammals reported a positive correlation between the two types of AEs (Xin et al. 2008; Ma et al. 2009). The present work reveals a much more complex, differentiated relationship thanks to the separate analysis of different functional domains of transcripts.

The structure of the correlations between different types of AEs revealed here suggests two opposite trends: 1) tight coupling between alternative transcription and AS and 2) preferential use of different AEs by two classes of genes with different dominant types of regulation. The genes in the first of these classes appear to be regulated, primarily, at the level of translation initiation, via variability of the 5′ UTR generated by ATI and, to a lesser extent, AS. By contrast, the genes

in the second class appear to be regulated primarily at the posttranslational level, via the formation of alternative protein forms resulting from AS in the CDS. Furthermore, our results suggest distinct roles for different types of AEs in the regulation of cellular processes and the possibility of common regulatory mechanisms for large groups of functionally related genes as demonstrated by the analysis of the distribution of different types of AE across the GO categories.

The coupling between ATI and AS in 5′ UTRs seems to receive a simple explanation from the requirement for optimal length (approximately 100 nucleotides, on average) of 5′ UTRs for efficient initiation of translation (Kozak 1978; Mignone et al. 2002): utilization of upstream transcription start sites necessitates AS to remove portions of the resulting long 5′ UTR (Lynch et al. 2005). The interdomain coupling between AS in the CDS and ATT in the 3′ UTR is more unexpected and suggests that the genes whose function is regulated through the formation of alternative protein forms are also regulated by the 3 ′UTRs, possibly, at the level of mRNA stability (Gallie 1991; Shyu et al. 2008). In addition, in eukaryotes, 3′ UTRs appear to contribute to the regulation of translation initiation via circularization of translated mRNAs (Hsu and Coca-Prados 1979; Komarova et al. 2006).

Low incidence of splicing in 3′ UTRs could be due to the high abundance of transcription termination signals in 3′ noncoding gene regions and also to the absence of strong constraint on the lengths of 3′ UTRs (Shabalina and Spiridonov 2004; Hong et al. 2006). For these reasons, termination of alternative transcripts with variable last coding exons may not require additional splicing, as it can be easily achieved at the next downstream transcription termination site. Splicing in 3′ UTRs might be also functionally unwarranted and avoided due to metabolic expenses associated with transcription of additional introns. Intron avoidance in 3′ UTRs together with the interdomain coupling between AS in the CDS and ATT in the 3′ UTR suggests coordinated regulation of these processes. Our results are in good agreement with the recently described connection between AS and alternative cleavage and polyadenylation across different tissues (Wang et al. 2008).

The present observations are compatible with the emerging understanding of the importance of cotranscriptional processing of mRNAs and functional connections between transcription initiation and splicing (Kornblihtt et al. 2004; Kornblihtt 2007). Numerous experimental results indicate that RNA polymerase II and transcription elongation factors recruit splicing factors to chromatin-associated "factories" in which transcription occurs concomitantly with various mRNA processing steps, including capping, splicing, cleavage/polyadenylation, and eventually, nucleocytosolic export (McCracken et al. 1997; Bentley 2002, 2005; Maniatis and Reed 2002; Hagiwara and Nojima 2007). The recruitment of splicing factors to the factories is specifically mediated by the phosphorylated, repetitive carboxy-terminal domain (CTD) of the RNA polymerase II largest subunit (Misteli and Spector 1999; Zeng and Berget 2000). Moreover, it has been shown that ultraviolet damage causes hyperphosphorylation of CTD with subsequent inhibition of transcription elongation and gene-specific modulation of the AS pattern that ultimately prevents apoptosis in irradiated cells (Munoz et al. 2009). The CTD has been shown to interact with splicing factors of the SR family and to directly regulate AS via exon skipping (de la Mata and Kornblihtt 2006). These results indicate that AS is not only coupled to transcription but is specifically regulated by the transcription machinery within the expression factories. Regulation of AS of specific genes in the factory critically depends on the structure of RNAP II promoter, providing direct evidence of coupling between transcription initiation and AS (Cramer et al. 1999).

In summary, the results of a genome-wide survey of AEs in mammalian transcripts suggest that alternative transcription is an even bigger source of mammalian transcriptome diversity than AS and that complex relationships between the two types of AEs, both synergistic and antagonistic, govern regulation of gene expression in mammals.

Supplementary Material

Supplementary figures S1–S3 and tables S1–S8 are available at *Genome Biology and Evolution* online (http://www.gbe.oxfordjournals.org/).

Acknowledgments

The research of S.A.S. and E.V.K. is supported by the intramural funds of the US Department of Health and Human Services (National Library of Medicine).

Literature Cited

Baek D, Davis C, Ewing B, Gordon D, Green P. 2007. Characterization and predictive discovery of evolutionarily conserved mammalian alternative promoters. Genome Res. 17:145–155.

Bentley D. 2002. The mRNA assembly line: transcription and processing machines in the same factory. Curr Opin Cell Biol. 14:336–342.

Bentley DL. 2005. Rules of engagement: co-transcriptional recruitment of pre-mRNA processing factors. Curr Opin Cell Biol. 17:251–256.

Blencowe BJ. 2006. Alternative splicing: new insights from global analyses. Cell. 126:37–47.

Cenik C, Derti A, Mellor JC, Berriz GF, Roth FP. 2010. Genome-wide functional analysis of human 5′ untranslated region introns. Genome Biol. 11:R29.

Cramer P, et al. 1999. Coupling of transcription with alternative splicing: RNA pol II promoters modulate SF2/ASF and 9G8 effects on an exonic splicing enhancer. Mol Cell. 4:251–258.

de la Mata M, Kornblihtt AR. 2006. RNA polymerase II C-terminal domain mediates regulation of alternative splicing by SRp20. Nat Struct Mol Biol. 13:973–980.

Gallie DR. 1991. The cap and poly(A) tail function synergistically to regulate mRNA translational efficiency. Genes Dev. 5:2108–2116.

Grillo G, et al. 2010. UTRdb and UTRsite (RELEASE 2010): a collection of sequences and regulatory motifs of the untranslated regions of eukaryotic mRNAs. Nucleic Acids Res. 38:D75–D80.
Hagiwara M, Nojima T. 2007. Cross-talks between transcription and post-transcriptional events within a "mRNA factory". J Biochem. 142:11–15.
Harris MA, et al. 2004. The Gene Ontology (GO) database and informatics resource. Nucleic Acids Res. 32:D258–D261.
Hong X, Scofield DG, Lynch M. 2006. Intron size, abundance, and distribution within untranslated regions of genes. Mol Biol Evol. 23:2392–2404.
Hsu MT, Coca-Prados M. 1979. Electron microscopic evidence for the circular form of RNA in the cytoplasm of eukaryotic cells. Nature. 280:339–340.
Kim E, Goren A, Ast G. 2008. Alternative splicing: current perspectives. Bioessays 30:38–47.
Komarova AV, Brocard M, Kean KM. 2006. The case for mRNA 5′ and 3′ end cross talk during translation in a eukaryotic cell. Prog Nucleic Acid Res Mol Biol. 81:331–367.
Kondrashov AS, Shabalina SA. 2002. Classification of common conserved sequences in mammalian intergenic regions. Hum Mol Genet. 11:669–674.
Kornblihtt AR. 2007. Coupling transcription and alternative splicing. Adv Exp Med Biol. 623:175–189.
Kornblihtt AR, de la Mata M, Fededa JP, Munoz MJ, Nogues G. 2004. Multiple links between transcription and splicing. RNA. 10: 1489–1498.
Kozak M. 1978. How do eukaryotic ribosomes select initiation regions in messenger RNA? Cell. 15:1109–1123.
Landry JR, Mager DL, Wilhelm BT. 2003. Complex controls: the role of alternative promoters in mammalian genomes. Trends Genet. 19:640–648.
Lynch M, Scofield DG, Hong X. 2005. The evolution of transcription-initiation sites. Mol Biol Evol. 22:1137–1146.
Ma X, et al. 2009. Systematic analysis of alternative promoters correlated with alternative splicing in human genes. Genomics 93:420–425.
Maniatis T, Reed R. 2002. An extensive network of coupling among gene expression machines. Nature. 416:499–506.
McCracken S, et al. 1997. The C-terminal domain of RNA polymerase II couples mRNA processing to transcription. Nature. 385:357–361.
Mignone F, Gissi C, Liuni S, Pesole G. 2002. Untranslated regions of mRNAs. Genome Biol. 3:REVIEWS0004.
Misteli T, Spector DL. 1999. RNA polymerase II targets pre-mRNA splicing factors to transcription sites in vivo. Mol Cell. 3:697–705.
Munoz MJ, et al. 2009. DNA damage regulates alternative splicing through inhibition of RNA polymerase II elongation. Cell. 137:708–720.
Ogurtsov AY, et al. 2008. Expression patterns of protein kinases correlate with gene architecture and evolutionary rates. PLoS One. 3:e3599.
Pan Q, Shai O, Lee LJ, Frey BJ, Blencowe BJ. 2008. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. Nat Genet. 40:1413–1415.
Resch A, et al. 2004. Assessing the impact of alternative splicing on domain interactions in the human proteome. J Proteome Res. 3:76–83.
Resch AM, Ogurtsov AY, Rogozin IB, Shabalina SA, Koonin EV. 2009. Evolution of alternative and constitutive regions of mammalian 5′ UTRs. BMC Genomics. 10:162.
Shabalina SA, et al. 2010. Distinct patterns of expression and evolution of intronless and intron-containing mammalian genes. Mol Biol Evol. 27:1745–1749.
Shabalina SA, Spiridonov NA. 2004. The mammalian transcriptome and the function of non-coding DNA sequences. Genome Biol. 5:105.
Shyu AB, Wilkinson MF, van Hoof A. 2008. Messenger RNA regulation: to translate or to degrade. Embo J. 27:471–481.
Sultan M, et al. 2008. A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome. Science. 321:956–960.
Wakaguri H, Yamashita R, Suzuki Y, Sugano S, Nakai K. 2008. DBTSS: database of transcription start sites, progress report 2008. Nucleic Acids Res. 36:D97–D101.
Wang ET, et al. 2008. Alternative isoform regulation in human tissue transcriptomes. Nature. 456:470–476.
Wegrzyn JL, Drudge TM, Valafar F, Hook V. 2008. Bioinformatic analyses of mammalian 5′-UTR sequence properties of mRNAs predicts alternative translation initiation sites. BMC Bioinformatics. 9:232.
Wilhelm BT, et al. 2008. Dynamic repertoire of a eukaryotic transcriptome surveyed at single-nucleotide resolution. Nature. 453:1239–1243.
Xin D, Hu L, Kong X. 2008. Alternative promoters influence alternative splicing at the genomic level. PLoS One. 3:e2377.
Xing Y, Lee C. 2006. Alternative splicing and RNA selection pressure–evolutionary consequences for eukaryotic genomes. Nat Rev Genet. 7:499–509.
Yamashita R, Wakaguri H, Sugano S, Suzuki Y, Nakai K. 2010. DBTSS provides a tissue specific dynamic view of transcription start sites. Nucleic Acids Res. 38:D98–D104.
Zeeberg BR, et al. 2003. GoMiner: a resource for biological interpretation of genomic and proteomic data. Genome Biol. 4:R28.
Zeng C, Berget SM. 2000. Participation of the C-terminal domain of RNA polymerase II in exon definition during pre-mRNA splicing. Mol Cell Biol. 20:8290–8301.

Associate editor: Kateryna Makova