Are all of the human exons alternatively spliced?

Feng-Chi Chen

Submitted: 24th January 2013; Received (in revised form): 25th March 2013

Abstract

Alternative mRNA splicing (AS) is a major mechanism for increasing regulatory complexity. A key concept in AS is the distinction between alternatively and constitutively spliced exons (ASEs and CSEs, respectively). ASEs and CSEs have been reported to be differentially regulated, and to have distinct biological properties. However, the recent flood of RNA-sequencing data has obscured the boundary between ASEs and CSEs. Researchers are beginning to question whether ‘authentic CSEs’ do exist, and whether the ASE/CSE distinction is biologically invalid. Here, I examine the influences of increasing transcriptome data on the human ASE/CSE classification and our past understanding of the properties of these two types of exons. Interestingly, although the percentage of human ASEs has increased dramatically in recent years, the overall distinction between ASEs and CSEs remain valid. For example, CSEs are longer, evolve more slowly, and less frequently correspond to intrinsically disordered protein regions than ASEs. In addition, only a relatively small number of human genes have their transcripts composed entirely of ASEs despite the large amount of high-throughput transcriptome information. Therefore, the ‘backbone’ concept of AS, in which CSEs constitute the invariant part and ASEs the flexible part of the transcript, appears to be generally true despite the increasing percentage of ASEs in the human exome.

Keywords: alternative mRNA splicing; human transcriptome; alternatively spliced exon; constitutively spliced exon; RNA sequencing

THE BIOLOGICAL SIGNIFICANCE OF EXON CLASSIFICATION IN ALTERNATIVE mRNA SPLICING

Pre-mRNA splicing occurs virtually in all of the known forms of life. This regulatory mechanism has been found in viruses [1, 2], prokaryotes [3, 4], fungi [5], plants [6] and animals [7]. In humans, alternative mRNA splicing (AS) is a major mechanism of increasing regulatory complexity [8], and is involved in biological regulations, development, genome evolution, and diseases [6, 7, 9–14].

One important concept in AS is the distinction between constitutively and alternatively spliced exons (CSEs and ASEs, respectively). CSEs are the exons that are included in all of the alternative transcript isoforms of a gene, whereas ASEs occur in only some of the isoforms. This distinction has important biological implications. Firstly, CSEs and ASEs are differentially regulated. Since CSEs are present in all of the transcript isoforms, the level and breadth of CSE expression are theoretically equivalent to those of the gene as a whole. By contrast, ASEs are regulated in a more sophisticated pattern so that they occur only in certain spatio-temporal conditions [8, 15–18]. Secondly, the regulability of ASE inclusion in the transcript conveys to the proteome a high level of biochemical flexibility, which may not be accomplished by CSEs alone. For example, the inclusion/exclusion of an ASE can significantly change the protein structure [19–21], protein–protein interaction networks [22, 23] or post-translational modifications [24]. All of these alterations can lead to significant changes in...
biological functions. Thirdly, the inclusion/exclusion of ASEs can cause reading frame shifts [25] or changes in the lengths/compositions of untranslated regions (UTRs). In the former case, AS may lead to significant changes in protein sequences or the initiation of nonsense-mediated decay [26, 27], thus adding another layer of regulatory complexity. In the latter case, the altered UTR sequences may cause changes in transcriptional or translational regulations [28–31]. Fourthly, CSEs and ASEs are known to have significantly different evolutionary rates [25, 32], and different paths of evolution [33]. Finally, the distinction between ASEs and CSEs brings us a concept that an mRNA transcript has a ‘backbone’ composed of CSEs, upon which ASEs can be added or removed according to biological conditions.

Whether all of the human exons are alternatively spliced is also an interesting topic in view of evolution. In prokaryotes and unicellular organisms, mRNA splicing rarely occurs. By contrast, in complex organisms, splicing is highly developed and plays important biological roles. In other words, in simple organisms, virtually all of the exons are CSEs. Yet the percentage and biological importance of ASEs increase dramatically in complex organisms. In humans, it has been reported that as many as 95% of the genes are alternatively spliced [34]. It is thus interesting to ask to what degree human, a model complex organism, has evolved away from the ‘backbone’ mechanism to adopt the ‘ad hoc assembly’ mechanism of mRNA splicing.

As well supported as it appears to be, the concept of ASE/CSE classification has been called into question thanks to the recent deluge of large-scale RNA-sequencing data. This is because the rapidly accumulating human transcriptome data provide evidence for previously unknown transcript isoforms, which have caused many CSEs to be re-classified as ASEs. Being aware of this frequent CSE-to-ASE changes, one may wonder whether all of the human exons will eventually be classified as ASEs, and whether all of the human transcripts are ad hoc assemblies of exons—i.e. combinations of ASEs that are dynamically assembled and disassembled according to biological conditions without a basic predefined framework. This idea is distinct from our previous ‘backbone’ concept of AS (in which CSEs constitute the ‘unchangeable backbone’ of a transcript), and will have important implications in biological regulations if proven true.

**THE COMPLEXITY IN THE ASE/CSE CLASSIFICATION**

The distinction between CSEs and ASEs appears to be simple. However, in complex organisms, this classification is sometimes ambiguous. This is because exons from different transcript isoforms usually partially overlap with one another. In Figure 1A, the transcript isoforms of the interested gene can be divided into two non-overlapping ‘exonic regions’. Note that each ‘exonic region’ contains partially or completely overlapping exons from different transcripts. For instance, in region #2, the exon from ENST00000221905 is longer than those from the other transcripts (partial overlap), whereas in region #3 all of the five exons are of the same length (complete overlap). Traditionally, a CSE is defined as an exon that is ‘always present’ in different transcript isoforms [12]. What is neglected in this definition is whether the exon ‘in its entirety’ should always present. For example, at region #1 in Figure 1B, both of the 5’ and 3’ boundaries of the exon change with isoforms. In such cases, it is difficult to distinguish between ‘CSEs’ and ‘ASEs’. Furthermore, consider region #5 in Figure 1A, although the exon is always present and the boundaries never change in different isoforms, this exon is part of the 5’UTR in one of the transcripts (ENST00000587433) but a coding exon in the others. In other words, this exon is a CSE in view of transcript structure, but can be viewed as an ASE when protein sequence is concerned. ASEs can be further classified into three groups—single (occurring only once in the isoforms), multiple (occurring at least twice with boundaries unchanged) and complex (occurring at least twice with boundaries changed; Figure 1A and B). Similarly, CSEs can be divided into simple and complex CSEs (Figure 1 and Supplementary Methods).

Adding to the complexity of the ASE/CSE classification is that different databases usually contain different numbers of transcript isoforms or transcripts of different lengths for the same genes. As shown in Figure 1C, for the human gene of interest (ENSG00000131149), three transcript isoforms are annotated by the ENSEMBL database (V69) to have a complete coding sequence and known protein product. However, the same gene is annotated by the UCSC (University of California Santa Cruz) Genome Browser (as of the November 2012 version) to have five such transcript isoforms. The inclusion of two short isoforms in the UCSC
Figure 1: Examples of ASE/CSE classification based on the known transcripts of ENSEMBL(V69)-annotated human genes (A) Rho GTPase activating protein 33 (ENSG00000000477); (B) Phosphatidylinositol glycan anchor biosynthesis, class V (ENSG00000060642); (C) KIAA0182 (ENSG0000003149) as compared with the corresponding transcripts retrieved from the UCSC Genome Browser (transcript IDs starting with 'uc'; as of the November 2012 version). Note that the transcripts analysed here are required to have complete coding sequences and known protein products. In (A) and (B), the empty and solid bars, respectively, represent untranslated regions and coding regions.
Figure 2: Statistics of the human exome and transcriptome according to the annotations of different ENSEMBL versions. (A) The percentages of different types of ASEs and the corresponding transcriptome information. The X-axis indicates the ENSEMBL version and the time of release. (B) The numbers of ASE-only genes. (C) The numbers of ASE/CSE exon type changes between different ENSEMBL versions. The X-axis indicates the ENSEMBL versions in the comparison; #CSE->ASE indicates the number of CSEs in the earlier version being redefined as ASEs in the latter version; similarly, #ASE->CSE indicates the number of ASEs redefined as CSEs. (D) The genic locations of CSE-to-ASE changes. Pure UTR (CDS): the exon of interest is part of the UTR (CDS) in all of the examined isoforms; Ambiguous: the exon includes both UTR and CDS, or the exon changes its role between UTR and CDS in different isoforms.
data significantly increases the number of ASEs while decreasing the number of CSEs.

**THE IMPACT OF HIGH-THROUGHPUT RNA-SEQUENCING DATA ON THE ASE/CSE CLASSIFICATION**

Recently, the number of large-scale RNA-sequencing datasets has been growing rapidly, adding a considerable number of previously unidentified transcript isoforms to the public databases. This flood of high-throughput RNA data has brought new information to the studies of mRNA splicing, and significantly increased the percentage of ASEs in the human exome. Here, I classify human exons into ASEs and CSEs (Figure 1) according to the ENSEMBL annotations (Supplementary Methods). As shown in Figure 2A, the percentage of ASEs in the human exome increased dramatically between Year 2008 and Year 2010, when high-throughput RNA-sequencing technologies became widely applicable [35]. In fact, the percentage of ASEs in the human exome has increased by ~2.5-fold for the past 8 years, from ~11.0% in ENSEMBL V24 (Year 2004) to >25% in ENSEMBL V69 (Year 2012) (Figure 2A). The average number of transcript isoforms per gene has increased by 65% during the same period (from ~1.42 to ~2.35 isoforms per gene). Of note, the percentage of simple ASEs has remained virtually invariant for the past 8 years, whereas those of multiple and complex ASEs have increased significantly, driving up the percentage of ASEs as a whole. This observation implies that a significant proportion of the newly added ASEs may be previously classified as CSEs (especially in the case of multiple ASEs, see Figure 1A). Figure 2A also indicates that the complexity of AS regulations has been underappreciated, for a considerable proportion of complex ASEs have just been emerging recently.

This rapid increase in the proportion of ASEs has caused many colleagues of the author to question whether this trend will continue until all human CSEs are reclassified as ASEs (i.e. will all human transcripts eventually be proven ‘ad hoc assembled’ without a basic framework?). However, the number of the ENSEMBL-annotated human genes that have their transcripts composed entirely of ASEs (‘ASE-only’ genes) has remained relatively small (<650; Figure 2B) despite the large amount of transcriptome information. The relatively small number of ASE-only genes does not seem to provide sufficient support for the generality of the ‘ad hoc assembly’ scenario. In addition, the percentage of ASEs appears to be slowly declining in the past 3 years (Figure 2A). Therefore, the current evidence appears to still support the ‘backbone’ scenario of AS.

One interesting question here is how the exon re-classifications occur. To address this issue, I conducted a case study on the comparison between ENSEMBL V61 and V63, where the number of ASEs changes dramatically (Figure 2C). The CSE-to-ASE changes in this comparison have three major causes: (i) the number of transcript isoforms for the same gene increases from V61 to V63 (Supplementary Figure S1A); (ii) the number of isoforms remain the same, but some exons annotated in V61 become absent in V63 for the same transcript (e.g. ENST0000396894 in Supplementary Figure S1B); and (iii) the number of transcript isoforms decreases in V63, but a new isoform (i.e. ENST00000543424) is added (Supplementary Figure S1C). Together these three reasons account for 97.0% of the CSE-to-ASE changes. Similarly, the vast majority (96.7%) of the ASE-to-CSE changes result from the decrease in the number of transcript isoforms (Supplementary Figure S2A), the presence of previously unannotated exons in V63 (Supplementary Figure S2B), or the decrease in exon number of a certain transcript despite the increase in isoform number in V63 (Supplementary Figure S2C). Close examination of Supplementary Figure S2A reveals that one of the transcripts (ENST00000455230) is annotated to be subject to nonsense-mediated decay. Interestingly, however, this known transcript is included because it actually has a known protein product according to ENSEMBL V61 (Supplementary Methods). Also noticeable in Supplementary Figure S2A is that three additional transcripts are annotated in V61 compared with in V63. According to the updated annotations (as of February 2013), these three transcripts have been merged into other transcripts. These observations suggest that many of the ASE/CSE re-classifications have resulted from ENSEMBL’s continuous efforts in correcting the annotations, and possibly also from the increasing availability of transcriptome information.

Notably, one important question in the ASE/CSE classification is which of the reported transcript isoforms are ‘biologically relevant’, and should be included for the analysis. Here, I used a series of
Figure 3: Re-examination of the results of previous AS-related studies using different versions of ENSEMBL human transcript annotations. (A) ASEs tend to correspond to intrinsically disordered regions; PIDR: proportion of amino acid residues predicted to fall within intrinsically disordered region. (B) CSEs and ASEs differ significantly in the $d_{SI}/d_5$-PIDR correlation (the $d_{SI}/d_5$ ratios were estimated by comparing human exons with the corresponding mouse exons). (C) ASEs tend to be shorter than CSEs. The statistical significance in (B) and (C), respectively, was evaluated by using the ANCOVA and the Wilcoxon Rank Sum test: **$p < .01$; ***$p < .001$. 
filters to ensure the quality and completeness of the analysed transcripts (Supplementary Methods). Nevertheless, we cannot rule out the possibility that some of the excluded isoforms are functionally relevant in certain biological conditions. Meanwhile, among the analysed transcripts, some yield significantly truncated proteins (e.g. ENST00000455230 among the analysed transcripts, some yield significantly truncated proteins) whose biological functions remain to be validated. The inclusion of such short isoforms for the ASE/CSE classification may be inappropriate. That said, we are still lacking a systematic and reliable approach to define the "functionally relevant" isoforms.

RE-EXAMINATION OF THE BIOLOGICAL PROPERTIES OF ASEs AND CSEs

Next, I ask whether the changes in the ASE/CSE classification in different ENSEMBL annotations might alter our understanding of the biological properties of ASEs and CSEs. Notably, the data sources and methods used here may differ from those used in previous AS-related studies. Therefore, a side-by-side comparison of the numerical results may be inappropriate.

One of the well-established properties in AS is that ASEs tend to correspond to intrinsically disordered protein regions (IDRs) [36]. To examine whether this proposition remains true, I calculated the proportion of amino acid residues that fell within IDR (PIDR) separately for ASEs and CSEs (Supplementary Methods) according to different versions of ENSEMBL annotations (V54, V59, V63 and V68). Interestingly, as shown in Figure 3A, the result is fairly consistent across different ENSEMBL versions, with ASEs having significantly higher PIDR than CSEs (all pairwise differences are statistically significant; \( P < .001 \) by Wilcoxon Rank Sum test). Therefore, changes in the ASE/CSE classification do not significantly affect this previously reported exon property.

A second known property in AS is the relaxed selection pressure and increased non-synonymous to synonymous substitution rate \((d_{n}/d_{s})\) ratio in ASEs as compared with in CSEs [25, 32, 37–40]. Of note, ASEs usually correspond to IDRs, which have also been reported to evolve rapidly [41–43]. This author’s team previously reported that the correlation between \(d_{n}/d_{s}\) ratio and PIDR differs significantly between CSEs and ASEs, although splicing and IDRs tended to correspond to the same genomic regions [40]. Here, I re-examine this proposition by applying the analysis of covariance (ANCOVA) [44] to datasets retrieved from different versions of ENSEMBL annotations using the same data filtering criteria as previously described [40]. With reference to our previous approach [40], the regression model is set to be

\[
d_{n}/d_{s} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_{12} X_1 X_2 + \varepsilon
\]

where \(X_1\) is the ASE/CSE exon type (CSE = 0; ASE = 1), \(X_2\) represents PIDR and \(\varepsilon\) is the error term.

Therefore, for CSEs,

\[
(d_{n}/d_{s}|X_1 = 0) = \beta_0 + \beta_2 X_2 + \varepsilon
\]

Whereas for ASEs,

\[
(d_{n}/d_{s}|X_1 = 1) = (\beta_0 + \beta_1) + (\beta_2 + \beta_{12}) X_2 + \varepsilon
\]

Therefore, if \(\beta_{12}\) deviates significantly from zero, the regression slopes for ASEs and CSEs are considered as different, and CSEs and ASEs are considered as having different \(d_{n}/d_{s}\)-PIDR correlations [44]. My results indicate that the CSE–ASE difference in the \(d_{n}/d_{s}\)-PIDR correlation holds well regardless of the changeable ASE/CSE classification across ENSEMBL versions (\(\beta_{12}\) deviates significantly from zero in all of the cases; Figure 3B).

The third property to be examined here is exon length. It has been recently reported that CSEs and ASEs had approximately the same lengths in the common ancestor of vertebrates. However, ASEs tend to shorten more rapidly than CSEs during vertebrate evolution. In other words, ASEs tend to be longer than ASEs in vertebrates generally holds well regardless of dataset changes (Supplementary Figure S3).

ARE ALL OF THE HUMAN EXONS ACTUALLY ASEs?

So far, I have shown that firstly, only a small fraction of human genes are ASE-only genes (Figure 2B) despite the significant increase in the percentage of
ASEs in recent years (Figure 2A). Secondly, the changing ASE/CSE classification has not affected the results of previous AS-related studies. Finally, the percentage of simple ASEs has remained fairly constant for years, yet the percentages of multiple and complex ASEs seem to have reached a plateau and begun to slowly decrease (Figure 2A).

The last observation implies that a large proportion of simple ASEs may have been discovered. However, for multiple and complex ASEs, it may be too early to jump to the same conclusion. A considerable number of currently identified CSEs are likely to be redefined as multiple or complex ASEs in the future. Yet, interestingly, ASEs can also be redefined as CSEs (Figure 2C). Therefore, the ASE/CSE re-classification is not simply a one-way avenue. One interesting question here is whether the CSE-to-ASE changes occur mainly in UTRs or in coding sequences (CDSs). In the case of UTRs, the ASE/CSE re-classification may be attributable at least in part to the uncertainty of transcriptional initiation/termination sites. Meanwhile, if the exon re-classification occurs mainly in CDS, the human proteome annotations are obviously incomplete (or incorrect), and are subject to further revisions or expansions. Interestingly, the largest proportion of CSE-to-ASE changes between ENSEMBL versions occur in pure CDSs (i.e. the exons that are part of the CDS in all of the isoforms; Figure 2D). And a significant proportion of the exon type changes occur at the CDS–UTR boundaries (the ‘ambiguous’ type of exons). Only a small proportion of them occur at pure UTRs.

In view of the aforementioned observations, I speculate that most of the currently defined human CSEs may be present in most, if not all, of the biological conditions. In other words, most of these ‘CSEs’ should be ‘broadly expressed’ (if not ‘ubiquitously expressed’) across multiple tissues and developmental stages. Therefore, the ‘backbone’ scenario of AS may be true in most of the biological conditions. However, we cannot exclude the possibility that on rare conditions, some of these ‘CSEs’ may be excluded. When these conditions and the corresponding transcripts are identified, the ‘CSEs’ will be redefined as multiple or complex ASEs.

PERSPECTIVE

With the increasing availability of large-scale RNA-sequencing data and the advances in bioinformatics tools, it is now feasible to explore whether the ‘exonic expression breadth’ or ‘tissue specificity of exonic expression’ is quantitatively correlated with the structural/functional properties, selection pressure and biological role of an exon. Notably, however, these measures have their own limitations. For instance, exonic expression breadth and average exonic expression level may change as the number of RNA-sequenced tissues increases. In addition, these measurements may differ between normal and diseased states. Therefore, such quantitative measurements would be more meaningful if the applicable biological conditions are well defined.

Because AS is highly developed in complex organisms as compared with simpler organisms, it is also interesting to investigate whether the aforementioned correlations differ between lineages, and how such differences may be related to functional divergences. In addition, the epigenetic features and dynamics of individual exons may be important for determining when and where an exon should be included, and eventually determine the biological role of AS. How the dynamics in epigenome is correlated with variations in splicing patterns remains underexplored. Finally, a systems view of exons, in which exons from different genes are co-regulated and spliced together to conduct specific functions [18], is particularly interesting and should be a focus of systems biology studies in the future.

SUPPLEMENTARY DATA

Supplementary data are available online at http://bib.oxfordjournals.org/.

Key Points

- The human ASE/CSE classification usually changes with datasets and available transcriptome information.
- Despite the frequent changes in the ASE/CSE classification in recent years, the overall biological properties, including length, evolutionary rates and correspondence to protein structural disorderliness, remain distinct between ASEs and CSEs.
- The percentage of ASEs in the human exome increases by ~2.5-fold for the past 8 years. However, only a relatively small number (<650) of human genes have their transcripts composed entirely of ASEs.
- In general, the ‘backbone’ scenario of mRNA splicing, in which CSEs constitute the basic framework and ASEs the flexible part of transcripts, appears to be true despite the increasing percentage of ASEs in the human exome.

Acknowledgements

I thank Dr Ben-Yang Liao for constructive discussions and Ms Chia-Lin Pan for technical assistance. I would also like to
thank three anonymous reviewers for their constructive comments.

**FUNDING**

National Science Council, Taiwan, under contract (NSC-101–2311-B–400–003); intramural funding of National Health Research Institutes (PH-102-PP06).

**References**


