Evolutionarily conserved coupling of transcription and alternative splicing in the EPB41 (protein 4.1R) and EPB41L3 (protein 4.1B) genes

Jeff S. Tan, Narla Mohandas, John G. Conboy

Life Sciences Division and Genomics Department, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA
Red Cell Physiology Group, New York Blood Center, New York, NY 10021, USA

Received 20 April 2005; accepted 11 August 2005
Available online 20 October 2005

Abstract

Recent studies have shown that transcription and alternative splicing can be mechanistically coupled. In the EPB41 (protein 4.1R) and EPB41L3 (protein 4.1B) genes, we showed previously that promoter/alternative first exon choice is coupled to downstream splicing events in exon 2. Here we demonstrate that this coupling is conserved among several vertebrate classes from fish to mammals. The EPB41 and EPB41L3 genes from fish, bird, amphibian, and mammal genomes exhibit shared features including alternative first exons and differential splice acceptors in exon 2. In all cases, the 5′-most exon (exon 1A) splices exclusively to a weaker internal acceptor site in exon 2, skipping a fragment designated as exon 2V. Conversely, alternative first exons 1B and 1C always splice to the stronger first acceptor site, retaining exon 2V. These correlations are independent of cell type or species of origin. Since exon 2V contains a translation initiation site, splice variants generate protein isoforms with distinct N-termini. We propose that these genes represent a physiologically relevant model system for mechanistic analysis of transcription-coupled alternative splicing.

Gene expression within an organism is orchestrated through a variety of cellular mechanisms, a major one being alternative splicing of pre-mRNAs. It is estimated that as many as 60% of human gene products undergo differential splicing, with many transcripts having multiple splicing patterns and some having thousands [1–3]. This RNA processing can lead to differences in exon composition at the 5′ end, within the internal coding regions, or at the 3′ end [4,5]. Since this regulated inclusion or exclusion of selected exons allows the synthesis of multiple protein isoforms that possess unique biological properties, polypeptide variants from the same gene can engage in vastly different cellular processes.

Many cis-acting regulatory sequences that strongly influence splicing efficiency have been identified in the RNA within or near regulated alternative exons [6]. In addition, a number of trans-acting splicing factors that interact at these sites to control splicing decisions have been characterized, and the balance between splicing enhancers and silencers is thought to play a key role in regulation. However, the manner in which these elements are spatially and temporally integrated with other RNA processing events in the nucleus is not well understood. An abundance of data supports the general hypothesis that downstream processes such as RNA splicing and polyadenylation are mechanistically coupled to upstream transcriptional events [7–10]. Recent studies have demonstrated, moreover, that transcription can also be coupled to alternative pre-mRNA splicing; i.e., the nature of the transcriptional promoter can influence the ability of known splicing factors to enhance the efficiency of inclusion of specific alternative exons [11,12]. In addition, coactivator proteins that enhance steroid hormone receptor transcription and alter its splicing have recently been reported, further reinforcing this concept [13].

In previous studies, we presented evidence of coupling between transcription and alternative splicing in the EPB41 and EPB41L3 genes (also known as protein 4.1R and 4.1B, respectively) of human and mouse [14,15]. Specifically, both EPB41 and EPB41L3 possess alternative first exons that are
mutually exclusive and that can splice to different splice acceptor sites in downstream exon 2. For example, EPB41 gene transcription is driven from three different promoters and produces three classes of transcripts with different first exons. These exons, termed 1A, 1B, and 1C, are distinguished by their distant location from the coding exons (72–100 kb), as well as their specificity in splicing to exon 2. Whereas exon 1A splices to exon 2 to exclude a small portion at the 5' end (termed exon 2'), transcripts with 1B and 1C include the small exon 2' region along with the main body of exon 2. As there is a start codon located in exon 2', this alternative splicing effects a switch in translation initiation, which results in different protein products: 80 kDa (1A) vs 135 kDa (1B and 1C) [14]. A precisely analogous pattern is also observed for EPB41L3, involving multiple far-upstream promoters/first exons and their differential splicing to alternative splice acceptor sites in exon 2 [15].

The goal of the current study was to explore the evolutionary conservation of putative transcription-coupled splicing in the protein 4.1 genes among various vertebrate classes. Here, we report that the N-terminus complexity of EPB41 and EPB41L3 genes has been preserved across vertebrate classes from fish to mammals. Moreover, the first exons in all of these organisms follow strictly defined splicing patterns to alternative acceptor sites in exon 2, identical to that observed in the human. These results indicate that the coupling between upstream promoters and downstream alternative splicing has been conserved for at least 500 million years and suggest that the 5' complexity of the EPB41/EPB41L3 genes plays a critical but underappreciated role in regulating expression of these genes encoding essential cytoskeletal adaptor proteins. These genes may offer a physiologically relevant model system for studying mechanisms of coupling between transcription and alternative splicing.

Results

Complex 5’ gene structure is phylogenetically conserved in the EPB41/EPB41L3 genes across several vertebrate classes

Previous studies have shown that the mouse Epb4.1 and Epb4.113 genes, like those in human, also possess alternative first exons [14,15]. In this study, we analyzed genetic databases and recent genome assembly data to examine phylogenetic conservation of EPB41 and EPB41L3 gene organization among other vertebrate classes. For mammals, comparative genomics revealed strong orthologs of EPB41 exons 1A and 1C and EPB41L3 exons 1A and 1B in the dog (Fig. 1), as well as in the chimp, rat, and cow (data not shown). The authenticity of these first exons is supported by their high homology to the

![Fig. 1. Multiple first exons are present in the EPB41/EPB41L3 genes in several vertebrate classes. (A) 5’ gene organization of EPB41 for the various species. At least two alternative first exons have been identified in the dog, frog, and chicken: one each from the 1A and 1B/C type. In zebrafish a good candidate 1B/C is indicated; a predicted exon 1A has also been found in a zebrafish transcript. However, since the zebrafish genome assembly is currently incomplete, its identity is uncertain at this time. Note that the alternative first exons are distinguished by their large separation from the downstream exon 2. (B) First exon organization for EPB41L3. Multiple first exons were also found for EPB41L3 from human to fish, except for chicken. The small numbers between the exons indicate distance in kilobases. 1A, 1B, and 1C refer to distinct types of exons based on their splicing pattern to exon 2/2'. AUG1 represents the start codon located in exon 2' found in all species examined.](image-url)
validated human exons (~70–90% identity), their location upstream of exon 2, and in many cases their representation in bona fide cDNAs from the various species (see below).

For other vertebrate classes such as birds, amphibians, and fish, the orthologous first exons could not be identified directly by genome comparison, presumably due to sequence divergence among these noncoding exons. Therefore, we used a different strategy for identifying their alternative first exons. The alternative approach for finding first exons involved retrieving species-specific EPB41 and EPB41L3 transcripts from the genetic databases, screening for clones that contain sequences extending upstream of exon 2, and then mapping the 5′-end sequences of these transcripts back to the cognitive genome assemblies. Using this method, we were able to identify alternative exons for both genes in nearly all of the nonmammalian vertebrates considered in this study, including EPB41 orthologs of chicken, frog, and zebrafish (1B/C only), as well as EPB41L3 orthologs of frog and zebrafish (Fig. 1). A candidate zebrafish EPB41 exon 1A has also been identified, but this assignment is tentative due to the fact that its sequence is not yet represented in the zebrafish genome assembly.

Based on these data, the overall organization of the 4.1 genes has been conserved for at least 500 million years back to the emergence of fish. An intriguing feature of this structure is the very long 5′ intronic distances that have been evolutionarily preserved. Although the zebrafish genome assembly is not yet complete for these large 4.1 genes, a consistent pattern emerges preserved. Although the zebrafish genome assembly is not yet complete for these large 4.1 genes, a consistent pattern emerges from other vertebrate classes in which intron length increases with time up to mammals. For example, whereas the distance from other vertebrate classes in which intron length increases with time up to mammals. For example, whereas the distance from EPB41 exon 1A to exon 2 is approximately 49 kb in the frog and 56 kb in the chicken, in the human genome these exons are separated by ~100 kb. Similarly, EPB41L3 exon 1A is 85 kb upstream of exon 2 in the frog, but ~141 kb distant from exon 2 in human.

Since in the dog, chicken, frog, and zebrafish we could not find both 1B and 1C orthologs for either EPB41 or EPB41L3, we organized the first exons of the vertebrates in this study into two distinct groups: the 1A type and the 1B/C type. The 1A type represents the farthest upstream (most 5′) first exon, and the 1B/C type represents the remaining first exon(s). This positional classification was also accompanied by functional correlates: exons 1A consistently exhibited a splicing pattern that was distinct from all exons 1B and 1C (see also below). Alternative first exons splice to different acceptor sites in exon 2

From earlier studies of EPB41 and EPB41L3 alternative first exons in the human and mouse, it was shown that these exons possess their own transcriptional promoters and that they splice differently to downstream acceptor sites in exon 2 [14,15]. This unique splicing paradigm suggested coupling between upstream transcription events and downstream alternative splicing events (Fig. 2A). For both EPB41 and EPB41L3, the alternative first exons can be classified into two distinct types based on their splicing to exon 2. One category, involving the most upstream exon (1A), splices to an internal acceptor site in exon 2 and excludes a small portion at the 5′ end (exon 2′). Since there is a start codon in exon 2′, this skipping leads to translation initiation much farther downstream in exon 4, using an AUG2 that is conserved in all species examined. In contrast, first exons of the 1B/C type splice to the first acceptor site in exon 2 so as to include exon 2′. As it leads to alternative start codon usage, this differential splicing creates protein isoforms.

To determine whether the strict coupling between first exon choice and downstream exon 2 splicing is also conserved among vertebrate species other than human and mouse, we undertook a survey of EPB41 and EPB41L3 transcripts in the GenBank database. Representative transcripts were analyzed not only from dog, chicken, frog, and zebrafish genomes (Figs. 2B and 2C), but also from orangutan, pig, rat, cow, and horse (not shown). This analysis demonstrated unequivocally that, for all of these organisms, the 5′-most exon (1A) always splices to the internal (distal) acceptor site in exon 2, while the succeeding first exon(s) of the 1B/C type always splices to the more proximal site. This is a very robust observation, since a combined total of 35 cDNAs of type 1A and 30 cDNAs of type 1B/C always obeyed this coupling rule, independent of species or tissue type.

Although both the proximal and the distal splice sites are used, these acceptors are very disparate in quality. Using an algorithm designed to estimate splice site strength (see Materials and methods), we found that the pyrimidine-rich proximal acceptor site, without exception, is the much stronger one for both EPB41 and EPB41L3 in all species tested (Fig. 3). The internal (distal) splice sites are consistently more purine-rich and score much lower for predicted acceptor site strength. For instance, the proximal site of EPB41 from human to fish is on average 86.6 ± 1.78; compared to the distal site’s average value of 70.9 ± 1.79. Discrepancy between the two sites is even more pronounced for EPB41L3. For this gene, the proximal site has an average splice value of 92.7 ± 1.82, while for the distal site, the value is 67.88 ± 1.21 (excluding the zebrafish datum, which is unusually low). The conclusion from these data is that cells must possess an active regulatory mechanism to facilitate the consistent and selective splicing of 1A-type exons across large introns to the weaker distal acceptor site in exon 2.

**Conservation of exon 2′ primary sequences**

Although the EPB41 and EPB41L3 first exons do not exhibit primary sequence conservation across vertebrate classes, in contrast the short exon 2′ region is highly conserved from human to zebrafish (Table 1). For EPB41, the length of exon 2′ from human to fish is essentially fixed (17 ± 1 nt), whereas exon 2′ in EPB41L3 varies from 21 nt (chicken and frog), to 27 nt (mouse, rat, cow, zebrafish), to 33 nt (human and dog). Sequence alignments revealed significant identity of exon 2′ between EPB41 and EPB41L3 among the species, especially in an 11-nt region spanning the translation start site. Here, there is a motif AYCATGACRAC that is
present in all of the species examined, with only very minor deviations. Functionally, this conserved sequence is required for translation initiation at ATG1; it could also potentially function in the selection of the distal splice site for exons 1A.

In the mouse and human genes for EPB41 and EPB41L3, skipping of exon 2' prevents translation of exon 2 sequences, leading to translation initiation at ATG2 downstream in exon 4. This results in synthesis of protein isoforms with shorter N-terminal domains [14,15]. Since the ATG2 in exon 4 is present and conserved in all the species examined (data not shown), this supports the notion that the same mechanism operates in the 4.1 genes of avians, amphibians, and fish.

**Discussion**

This study demonstrates that complex S′ gene structure is phylogenetically conserved across several vertebrate classes for the genes encoding two members of the protein 4.1 family of cytoskeletal proteins: EPB41 and EPB41L3. For these genes from fish to mammals, genome and transcript analyses reveal multiple alternative first exons that map far upstream of the coding regions, as well as unique splicing patterns for these exons to specific splice acceptor sites in exon 2 that result in differential translation initiation. The most remarkable finding is that, despite the long intronic sequences that separate these first
exons and their downstream splicing acceptors, there is a strictly
obeyed paradigm: transcripts that initiate at exon 1A always skip
exon 2’, while transcripts that start at exons 1B/C always include
exon 2’ (Fig. 4). This specific splicing is observed across many
tissue sources and in diverse vertebrate classes ranging from fish
to mammals (this study,[14]).

Since exons 1A and 1B/C have been associated with
independent transcriptional promoters, it is assumed that the
splicing correlations of 4.1 transcripts are due to coupling
between transcription and alternative splicing. Promoter
activity has been demonstrated for flanking, upstream
sequences of human EPB41 exons 1A and 1B, as well as
EPB41L3 exons 1A and 1B, using luciferase as a reporter
[14,15]. Since other mammalian exons in this study share
significant sequence identity with the human orthologs, this
strongly suggests their authenticity as being true first exons.
For the nonmammalian vertebrates, the strong similarity in
genomic organization and splicing patterns to exon 2
suggests these exons as being functionally homologous first
exons with associated promoter activity. In addition, analysis
of 5’ sequences in dozens of 4.1 cDNAs from multiple
vertebrate species revealed no evidence of any exons
upstream of those designated here as first exons.

An intriguing issue surrounding these first exons is their
ability to discriminate, from a position tens of kilobases
upstream of exon 2, between splice acceptor sites that are
only 17–33 nt apart. Computational analyses indicate that the
proximal and distal acceptor sites exhibit dramatic differences
in intrinsic splice site “strength”—with the former always
stronger than the latter—suggesting that the first acceptor
would likely be the default splice site. Future mechanistic
studies will need to explain how exon 1A can selectively
bypass the stronger acceptor to choose the apparently weaker
site just downstream.

One attractive hypothesis is that there is a physical and
functional coupling between transcription and alternative
splicing [11–13,18]. Since the long CTD repeat domain of
the RNA Pol II large subunit has been proposed as a platform
for assembly of protein complexes required for downstream
RNA processing events, it may also serve a similar function
for proteins that mediate alternative splicing decisions [19–
23]. A number of interesting proteins that can interact with
both CTD and RNA have been identified as potential
candidates for coupling of transcription and splicing. Among
these are SAF-B, SCAFs (SR-like CTD-associated factors),
TLS/FUS-related proteins, p54nrb/PSF, and U2AF65-related
proteins CAPERα and CAPERβ [24–30,13,18]. But how
might promoter structure influence alternative splicing? If
loading of splicing cofactors onto the CTD is influenced by
promoter-specific regulatory elements, then it is easy to
speculate that the promoters associated with each alternative
first exon might assemble different cofactors that determine
the splicing pattern at exon 2’. The observed ability of
CAPERα and CAPERβ to modulate downstream alternative
splicing specifically from steroid hormone-responsive, but not
nonresponsive, promoters represents a precedent for this type
of model [13].

Alternatively, the promoter might influence downstream
splicing events in a couple of other ways. The promoter may
recruit molecules that can phosphorylate the CTD. Since
phosphorylated CTD has been shown to be much stronger in
activating the spliceosomal assembly than its nonphosphory-
lated form, molecules that alter the level of CTD phosphor-
ylation can have a potent impact in splicing regulation [31].
Finally, promoters might also influence splicing events by
recruiting factors that can manipulate transcription kinetics. It
has been reported that introduction of transcriptional pause
sites downstream of a weak splice site can improve its
efficiency of recognition by the spliceosomal machinery
[32,33]. However, this particular model is difficult to
reconcile for the regulated splicing events at EPB41 and
EPB41L3 exon 2’. Since the stronger splice acceptor site is
transcribed before the weaker competing site, it is not
apparent how transcriptional pausing could impart a kinetic
advantage for the weaker site.

Regardless of its exact mechanism, the differential
splicing of alternative exons to acceptor sites in their
downstream exon 2 has important functional ramifications
Table 1
Conservation of exon 2’ in EPB41 and EPB41L3

<table>
<thead>
<tr>
<th>Species</th>
<th>EPB41 Exon 2’/2</th>
<th>EPB41L3 Exon 2’/2</th>
</tr>
</thead>
<tbody>
<tr>
<td>human</td>
<td>CAACATCATGAGCAACAG</td>
<td>ACAACTACAGCCACAGC</td>
</tr>
<tr>
<td>mouse</td>
<td>GAACATACAGGACACAG</td>
<td>GAGAACACACAGCCACAG</td>
</tr>
<tr>
<td>dog</td>
<td>CAACATCATGAGCAACAG</td>
<td>CAACATACAGCCACAGC</td>
</tr>
<tr>
<td>chicken</td>
<td>CACCACTGAGCAACAGC</td>
<td>GAGAACACACAGCCACAG</td>
</tr>
<tr>
<td>frog</td>
<td>GTATACATACAGCCACAG</td>
<td>TGATACATACAGCCACAG</td>
</tr>
<tr>
<td>zebrafish</td>
<td>TCCTGACCAGACAGC</td>
<td>TCCTGACCAGACAGC</td>
</tr>
</tbody>
</table>

The nucleotide sequences of exon 2’ in both (A) EPB41 and (B) EPB41L3 for all the species examined are shown here. There is an 11-nt sequence, marked by the motif AYCAGACRAC (highlighted box), which is conserved for both EPB41 and EPB41L3 among all the species. This stretch of sequence may play an important role in the selection of the distal splice site for exon type 1A. The start codon in exon 2’ is highlighted in red.

of the encoded protein 4.1 polypeptides. By regulating expression of the translation initiation site in exon 2’, upstream transcriptional events ultimately regulate the synthesis of protein isoforms with very different N-terminal domains. Since the N-terminal domain in the longer isoform of EPB41 (135 kDa) can significantly alter the binding of EPB41 to membrane proteins such as band 3 and glycophorin C, transcription-splicing regulation may be an essential determinant of important biological properties of the protein [34]. The observation that this mechanism has been extant for many millions of years of evolution further supports its biological importance as a method of gene regulation and expression.

Finally, we speculate that coupling between transcription and alternative splicing may be far more widespread than is currently appreciated. The clear manifestation of this phenomenon in the 4.1 genes is due to involvement of independent first exons with completely distinct nucleotide sequences represented in the mature mRNAs. It is entirely possible that a similar coupling occurs in many genes in which no such “sequence tag” remains at the 5’ end of the mRNA. Precedence for such a model has been reported using synthetic constructs to show that an alternative exon-containing minigene can exhibit different splicing efficiencies when transcriptionally activated by different promoters [11,12]. An analogous situation could apply to natural endogenous genes by assembly of functionally distinct transcription complexes, involving alternative enhancer cofactors, without overt changes in the 5’ end of the transcribed RNA. In such cases the transcription-alternative splicing connection would be impossible to discern from mRNA analysis alone. Future studies will be needed to investigate this hypothesis.

Materials and methods

Identification of alternative EPB41/EPB41L3 5’ exons in various vertebrate classes

Alternative 5’ exons of the human and mouse EPB41/EPB41L3 genes were identified in earlier studies through genetic database searches (GenBank and the UCSC genomics Web site, http://www.genome.ucsc.edu) and RACE experiments [14,15]. To identify alternative 5’ exons in the EPB41/EPB41L3 orthologs of the dog (Canis familiaris), chicken (Gallus gallus), frog (Xenopus tropicalis), and zebrafish (Danio rerio), the exon 2 sequences of human EPB41 and EPB41L3 were first used in BLAT searches in the UCSC genomics database to identify their orthologs in these species. Once the orthologs were identified, alternative 5’ sequences for this gene were analyzed through identification of existing, known clones. In cases for which no submitted data were available, putative first exons were deduced through homology to the human first exon (1A, 1B, or 1C) sequences; sequences that showed significant identity to the human exons and that mapped upstream of exon 2 in the expected locations were accepted as orthologs. This approach led to the identification of multiple first exons for both EPB41 and EPB41L3 in almost all of the species analyzed. Note that the zebrafish EPB41 clone extends the previously reported 5’ sequence so as to include the bona fide translation initiation site AUG1, which encodes an N-terminal tetrapeptide MTTD, highly homologous to the mammalian MTTE N-terminal motif [16].

Analysis of splice site strength in exon 2’ of EPB41/EPB41L3

Splice site strengths of both the proximal and the distal acceptors in exon 2’ were computed using a special program (http://www.genet.sickkids.on.ca/~ali/splicesitescore.html) [17]. To compute the acceptor scores, a 15-bp sequence was inputted into the program; this sequence consisted of 14 bp before and 1 bp after the splice junction. Although this program is designed...
for human genes, it was used to assess the splice acceptors for all the species considered in this study; since splice sites are well conserved through evolution, it is assumed that the program would also be applicable to nonhuman species.

Acknowledgments

The authors are grateful to Philippe Gascard and Sharon Krauss for helpful discussions. This research was supported by National Institutes of Health Grants DK32904 and HL45182 and by the Director, Office of Biological and Environmental Research, U.S. Department of Energy under Contract DE-AC03-76SF00098.

References