

## Assessing the Impact of Alternative Splicing on Domain Interactions in the Human Proteome

Alissa Resch, Yi Xing, Barmak Modrek, Michael Gorlick, Robert Riley, and Christopher Lee\*

*Molecular Biology Institute, Center for Genomics and Proteomics, and Department of Chemistry & Biochemistry, University of California—Los Angeles, Los Angeles, California 90095-1570*

Received August 18, 2003

We have constructed a database of alternatively spliced protein forms (ASP), consisting of 13 384 protein isoform sequences of 4422 human genes ([www.bioinformatics.ucla.edu/ASP](http://www.bioinformatics.ucla.edu/ASP)). We identified fifty protein domain types that were selectively removed by alternative splicing at much higher frequencies than average ( $p$ -value < 0.01). These include many well-known protein-interaction domains (e.g., KRAB; ankyrin repeats; Kelch) including some that have been previously shown to be regulated functionally by alternative splicing (e.g., collagen domain). We present a number of novel examples (Kruppel transcription factors; Pbx2; Enc1) from the ASP database, illustrating how this pattern of alternative splicing changes the structure of a biological pathway, by redirecting protein interaction networks at key switch points. Our bioinformatics analysis indicates that a major impact of alternative splicing is removal of protein–protein interaction domains that mediate key linkages in protein interaction networks. ASP expands the available dataset of human alternatively spliced protein forms from 1989 human genes (SwissProt release 42) to 5413 (nonredundant set, ASP + SwissProt), a nearly 3-fold increase. ASP will enhance the existing pool of protein sequences that are searched by mass spectroscopy software during the identification of peptide fragments.

**Keywords:** alternative splicing • human proteome • protein interaction domains • bioinformatics

### Introduction

Alternative splicing has emerged as a major mechanism for expanding and regulating the repertoire of gene function. Previously considered to be an unusual event (estimated to occur in 5–15% of genes), it has recently been identified in 30–60% of human genes by large-scale genomics studies.<sup>1–6</sup> For example, approximately 30 000 alternative splices have been detected in the human genome alone,<sup>7</sup> based on mapping of expressed sequences (experimental mRNAs and ESTs) that show different exon-intron splicing patterns when aligned to the human genomic sequence. The *Drosophila Dscam* gene provides an extreme example: combinatorial alternative splicing of four “exon cassettes” is capable of generating up to 38 016 distinct protein isoforms from this single gene.<sup>8</sup>

Alternative splice data evidently have the potential to contribute substantially to our understanding of proteomic diversity and function. However, although most alternative splicing is studied at the nucleic acid level, many functional questions can only be answered by analyzing the protein products. Recently, studies of SwissProt and RefSeq annotations have found two connections between alternative splicing and protein functional impact. First, alternative splicing events that remove complete protein domains are more frequent than expected by random chance;<sup>9</sup> second, alternative splicing shows a statistically significant bias toward occurring in proteins with

specific domains or Gene Ontology annotations.<sup>10</sup> These data suggest that such alternative splicing events are functional and raise the interesting question of what their detailed effects might be. Ordinarily, it has been difficult for biologists to go from a particular alternative splice identified in genomics data, to a reliable protein isoform sequence and its functional domain impact. Only a small fraction of the alternative splices detected in the human genome<sup>7</sup> is available in proteomics databases.

To make this connection, we have constructed a database of 13 384 protein isoforms generated by alternative splicing, based on a well-documented bioinformatics approach.<sup>6,7,11–13</sup> In addition to being useful for many biologists interested in finding possible splice forms for a particular protein, this “Alternatively Spliced Protein forms” (ASP) database provides a proteome-wide sample of the functional impact of alternative splicing. We have analyzed the effects of alternative splicing on the removal or modulation of protein functional domains and binding sites. We have also sought to validate the ASP data by comparison with independent literature and with alternative splicing data in the human-curated SwissProt database. The ASP data indicate that a ubiquitous function of alternative splicing is modulating protein interaction networks, by adding or removing protein–protein interaction domains that are key linkages.

### Methods

**Generation of Protein Isoforms.** We identified alternative splicing events in human genes as previously described,<sup>6</sup> using

\* To whom correspondence should be addressed. Tel: (310) 825-7374. Fax: (310) 267-0248. E-mail: leec@mbi.ucla.edu.

NCBI's January 2002 draft human genomic sequence<sup>4</sup> and UniGene human EST data<sup>14</sup> downloaded in January 2002. Unlike mRNA sequences, which produce full-length transcripts that average 2209 nucleotides in length, ESTs are short, fragmentary expressed sequences that average 506 nucleotides in length. We assembled a mixture of mRNA and fragmentary EST data to generate full-length mRNA and protein isoforms for our alternatively spliced genes. Individual exons were identified by their start and end positions in the genomic sequence; individual splices were identified by their 5' and 3' splice-site positions in the genomic sequence. A transcript was defined as a list of exons and a list of associated splices connecting them. The major transcript was identified by counting the number of mRNAs and ESTs that support each transcript (i.e., the exons and splices observed in the EST must be consistent with the transcript). The longest open reading frame was identified in each candidate transcript.

The set of transcripts was filtered by a variety of criteria. (1) Only major–minor isoform pairs resulting in a change to the protein product were retained. (2) No transcripts incorporating nonconsensus splice sites were permitted. (3) The transcript's longest ORF must be full length, i.e., begin with AUG and end with a stop codon. (4) Minor form protein products must have at least 50% identity to the major form (that is, no more than half of the major form protein sequence can be changed or removed), and a minimum length of 50 amino acids. This produced a subset of 13 384 protein isoforms in 4422 human genes. Full details and validation results of the isoform generation procedure have been presented elsewhere.<sup>12,13</sup>

**Functional Annotation of Protein Isoforms.** We used RPS-BLAST to search our protein isoforms for matches to the Conserved Domain Database (CDD)<sup>15</sup> of Pfam<sup>16</sup> and SMART<sup>17</sup> domain sequences. We considered hits only with expectation values less than  $10^{-4}$ . A total of 8674 domain hits were found in the ASP isoforms, representing 883 distinct domain types. The intervals of alignment of the domain hits were recorded; hits that matched less than 70% of the original CDD domain sequence were discarded. Next, we aligned the major–minor protein isoform pairs, and recorded the intervals in each that were removed or replaced in the other isoform. We then mapped the domain intervals versus the removal intervals on each isoform sequence, to identify domains that were affected by alternative splicing. A total of 554 distinct domain hits were found to overlap the intervals removed by alternative splicing.

## Results

**ASP Alternatively Spliced Protein forms Database.** As the initial basis for the ASP database, we constructed a filtered set of 13 384 alternatively spliced protein sequences. To do so, we first identified the major transcript isoform for each gene (expressed as a sequence of exon IDs), defined as the isoform supported by the largest number of mRNA and EST sequences. We also generated alternatively spliced transcript isoform sequences (which we will refer to as “minor isoforms”), and filtered all the isoforms by a series of criteria for both their splice sites and protein translation (see the Methods section). This produced a reduced set of 4422 major isoforms and 8962 minor isoforms with full-length protein sequences. This dataset was designed as a highly filtered subset of “productive” alternative splice forms (i.e., alternative splices which result in a functional protein product different from the major protein isoform). The purpose of ASP is to give a large and representative sample of the effects of alternative splicing on the pro-

teome. It is important to emphasize that many correct, authentic protein isoforms may have been excluded by our filters.

How does alternative splicing modify protein domain architecture? Using the PFAM<sup>16</sup> and SMART<sup>17</sup> domain databases, we identified a total of 554 domains in ASP that were modified by an alternative splicing event. In 509 cases (92%), alternative splicing removed residues from a domain identified in the major isoform, producing a minor isoform that lacked part or all of a domain. In 73% of the cases, alternative splicing removed more than 30% of a domain, which we will define throughout this paper as “domain disruption”, because removal of such a large proportion of a domain is likely to disrupt its function and structure. The most striking feature of the domain removal statistics is that about half remove essentially the entire domain. Alternative splicing events removing a small portion of a domain occurred at a much lower, relatively uniform level, with a slightly higher rate (11%) for very small removals (less than a tenth of the domain, typically just a few residues). In a small fraction of cases (8%), alternative splicing inserted new domain residues into a minor isoform. In most of these cases, this inserted an entirely new domain that was lacking from the major isoform.

**Identification of Protein Domains Removed Preferentially by Alternative Splicing.** To find protein domains that are preferentially removed by alternative splicing, we counted the total number of times a removal occurred in each domain, vs the total number of occurrences of each domain. Of the total number of all domains (4862 in our filtered major isoform set), we observed 509 cases in which a domain was either completely or partially deleted from the major isoform via an alternative splicing event, yielding a 10.5% average probability of a domain being removed by alternative splicing. This value is the product of three factors: the fraction of alternative splices that are in coding regions; the fraction of alternative splice events that cause removal of residues from the major isoform; and the fraction of the protein sequence that is removed by each splice.

By contrast, some specific domains are alternatively spliced much more frequently. For example, out of 24 total occurrences of the annexin domain, 13 were alternatively spliced (54%). To assess the statistical significance of these results, we calculated the log odds ratio for the null hypothesis that the alternative splicing frequency of a domain is the same as that for all domains. For the annexin domain, the LOD score was 7.5 ( $p = 10^{-7.5}$ ).

Fifty domains scored above LOD 2 ( $p$ -value  $< 0.01$ ; Table 1). To control for varying domain length, we also calculated the frequency of removal for each domain normalized by its length, and a LOD score for its deviation from the average normalized frequency for all domains (NLOD in Table 1). Forty-eight of the fifty domains (96%) also had a NLOD  $> 2$ , indicating that these results are not artifacts of domain length. Our analysis detected several domains that are known to be targets of alternative splicing. For example, alternative splicing of the collagen domain (27% in our dataset; LOD score 4.1) has been shown to play an important role in modulating the function of the extracellular matrix in different tissues.<sup>18–20</sup> We detected unusually high levels of alternative splicing in many protein–protein interaction domains, including the KRAB domain (57%, LOD 7.4), ankyrin repeat (ank, 33%, LOD 4.5), and Kelch domain (31%, LOD 3.4). Several DNA-binding domains also displayed unusually high alternative splicing levels, including C4-type Zinc fingers (zf-C4, 53%, LOD 5.9), SANT domain

**Table 1.** Preferential Alternative Splicing of Specific Protein Domains<sup>a</sup>

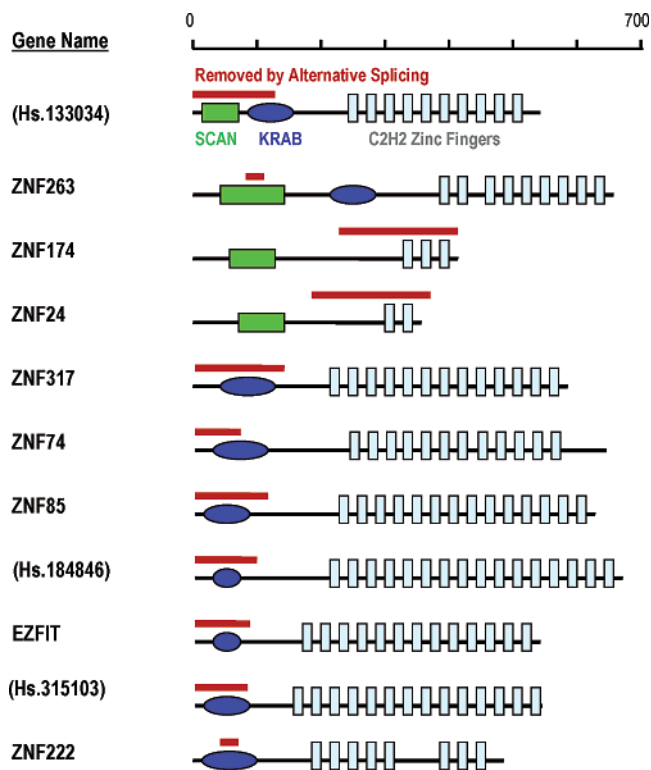
CDD_id	domain	no. removed	no. total	% removed	LOD	NLOD
pfam00191	annexin	13	24	54.2	7.5	7.9
smart00349	KRAB	12	21	57.1	7.4	10.7
pfam00105	zf-C4	10	19	52.6	5.9	7.3
pfam02946	GTF2I	4	4	100.0	4.9	7.8
pfam00108	thiolase	4	4	100.0	4.9	3.6
pfam00382	transcript_fac2	4	4	100.0	4.9	6.5
pfam00023	ank	13	39	33.3	4.5	11.1
pfam01391	Collagen	18	68	26.5	4.1	8.7
smart00395	SANT	5	8	62.5	4.1	5.9
pfam00646	F-box	3	3	100.0	3.9	9.7
pfam01585	G-patch	3	3	100.0	3.9	5.6
pfam03020	LEM	3	3	100.0	3.9	6.8
smart00290	ZnF_UBP	3	3	100.0	3.9	6.2
pfam00790	VHS	3	3	100.0	3.9	2.0
pfam02198	SAM_PNT	3	3	100.0	3.9	4.2
smart00501	BRIGHT	3	3	100.0	3.9	4.0
smart00612	Kelch	10	32	31.3	3.4	7.1
smart00131	KU	5	10	50.0	3.4	5.2
pfam00605	IRF	3	4	75.0	3.3	3.2
pfam01505	Vault	3	4	75.0	3.3	4.4
pfam02771	Acyl-CoA_dh_N	2	2	100.0	2.9	7.1
smart00562	NDK	2	2	100.0	2.9	6.7
pfam00217	ATP-gua_Ptrans	2	2	100.0	2.9	3.0
pfam00364	biotin_lipoyl	2	2	100.0	2.9	5.0
pfam02776	TPP_enzymes_N	2	2	100.0	2.9	3.7
pfam00326	Peptidase_S9	2	2	100.0	2.9	3.5
pfam00418	tubulin-binding	2	2	100.0	2.9	4.4
pfam00970	FAD_binding_6	2	2	100.0	2.9	3.1
pfam00804	Syntaxin	3	5	60.0	2.8	1.7
pfam00046	homeobox	5	13	38.5	2.7	4.3
pfam00412	LIM	8	28	28.6	2.6	5.1
pfam00271	helicase_C	8	29	27.6	2.5	3.9
smart00338	BRLZ	4	10	40.0	2.5	3.7
pfam00179	UQ_con	3	6	50.0	2.5	2.2
pfam00620	RhoGAP	3	6	50.0	2.5	2.1
pfam00627	UBA	3	6	50.0	2.5	3.3
pfam00641	zf-RanBP	3	6	50.0	2.5	4.6
smart00385	CYCLIN	3	6	50.0	2.5	3.0
pfam00123	hormone2	2	3	66.7	2.4	4.1
pfam00774	Ca_channel_B	2	3	66.7	2.4	2.0
pfam01335	DED	2	3	66.7	2.4	2.9
pfam02760	HIN	2	3	66.7	2.4	2.1
smart00027	EH	3	7	42.9	2.2	2.7
smart00175	RAB	7	28	25.0	2.1	1.6
pfam00071	ras	7	29	24.1	2	2.6
pfam00415	RCC1	3	8	37.5	2	3.3
pfam00432	prenyltrans	2	4	50.0	2	3.3
pfam00788	RA	2	4	50.0	2	2.5
smart00398	HMG	2	4	50.0	2	2.9

<sup>a</sup> Fifty protein domains were observed to be removed by alternative splicing more frequently than average, with a p-value < 0.01. The table lists the number of times the domain was removed by alternative splicing (no. removed), the total number of occurrences of the domain in ASP major isoforms (no. total), log-odds ratio for statistical significance as a function of total counts of the domain (LOD) and as a function of the total number of amino acids summed for all occurrences of the domain (NLOD; see text).

(SANT, 63%, LOD 4.1), and homeobox domain (homeobox, 39%, LOD 2.7).

How do these domain splicing events regulate biological function? Specifically, how does alternative splicing impact protein interaction networks and the signaling pathways that they form? To address these questions, we present several detailed case studies of alternative splicing of protein domains identified by ASP and propose pathway switching mechanisms for several of the novel alternative splice forms discovered in our analysis.

**Preferential Removal of Protein Interaction Domains in Kruppel Family Transcription Factors.** We have analyzed alternative domain splicing in the well-studied Kruppel family

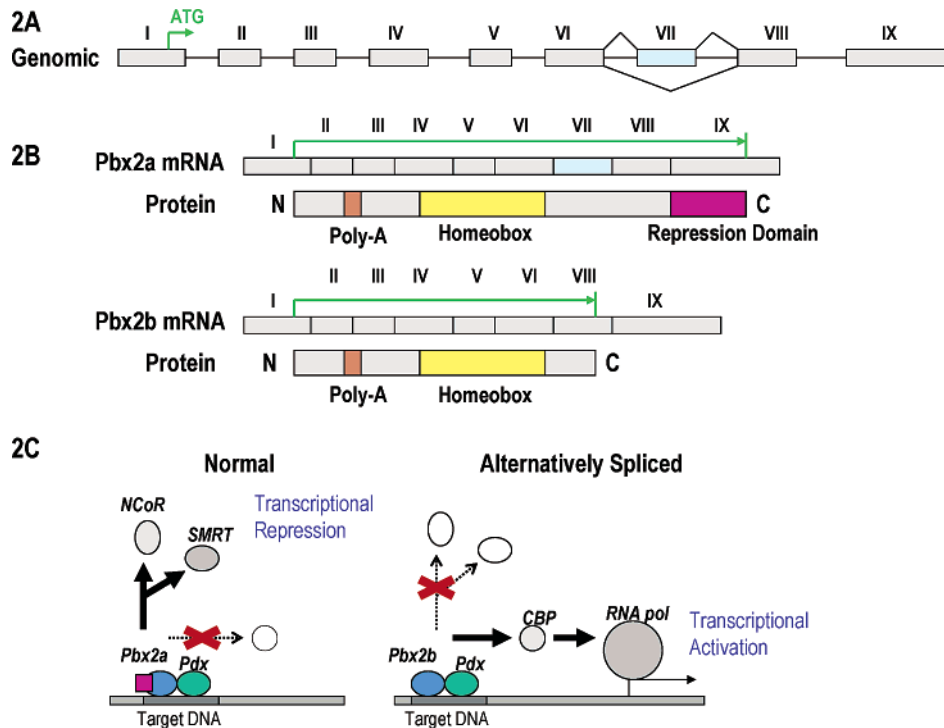


**Figure 1.** Preferential removal of protein interaction domains in Kruppel family transcription factors. SCAN dimerization domains (green), KRAB repressor domains (blue), and C2H2-zinc finger DNA-binding domains (light blue) are indicated, and the amino acids removed by alternative splicing are covered by a red bar. Proteins are named by their gene symbol, or UniGene cluster ID (in parentheses) for novel genes.

of transcription factors, which each contain at least one protein-interaction domain (KRAB repressor domain and/or SCAN dimerization domain) and multiple DNA-binding domains (C2H2-type zinc finger domains). We searched our ASP database for Kruppel family members, and identified eleven alternatively spliced genes (Figure 1). Four family members also contained a SCAN domain, a leucine-rich protein interaction domain thought to mediate homo- and heterodimerization events between specific members of the SCAN family of transcription factors.<sup>21</sup> In nine genes, alternative splicing preferentially removed the protein interaction domain (KRAB or SCAN) via in-frame deletions, while removing zinc fingers in only two cases. In all nine of the proteins, the size of the removal constituted domain disruption of the protein interaction domain.

This is a statistically significant result. Given the counts of domains that are alternatively spliced in this family (10 out of the 13 total protein interaction domains, vs only 5 of the total 114 DNA binding domains), the p-value for the null hypothesis that both types of domains are equally affected by alternative splicing is 10<sup>-9.3</sup>. The correspondence between the interval of amino acids deleted by alternative splicing, and the protein-interaction domain boundaries, is very striking within this specific family.

Our analysis of the new isoforms predicts that they will bind to the same target sequences, but exert a positive rather than negative effect due to loss of the repression domain. This mechanism for modulating the action of a transcription factor is well-documented in both mutation studies and in the



**Figure 2.** Alternative splicing of the Pbx2 transcription factor. (A) Genomic structure of the *PBX2* gene. Exons are shown as gray boxes and the alternatively spliced exon is colored blue. The truncated isoform *Pbx2b* is generated by an out-of-frame exon skip involving exon VII. (B) The two alternative protein isoforms of *PBX2* found in ASP. Alternative splicing removes the transcription repression domain from isoform *Pbx2b*. The protein-coding region for each isoform is represented as a green arrow. (C) Proposed model of pathway switching by alternative splicing. Removal of the *Pbx2* transcription repression domain provides a switching mechanism for regulating gene expression. The *Pbx2* protein (blue) forms a heterodimer with the *Pdx* protein (green), a member of the Hox family of transcription factors. The C-terminal transcription repression domain of *Pbx2a* (magenta rectangle) is thought to recruit the co-repressor proteins NCoR and SMRT (represented by gray ovals), thereby repressing transcription of the target DNA sequence. Removal of the C-terminal repression domain from *Pbx2b* blocks recruitment of the co-repressor proteins, while simultaneously enabling the *Pdx* protein (part of the *Pbx2b-Pdx* heterodimer) to recruit the co-activator CBP and RNA polymerase (gray ovals), activating transcription of the target DNA sequence.

evolutionary history of transcription factor families.<sup>22,23</sup> Indeed, the common domain architectures among evolutionary homologues of the Kruppel family follow the same pattern of variation (removal of one or both of the protein interaction domains; see the Supporting Information). Evolution has apparently employed this kind of domain disruption and addition to create useful new functions, supporting the suggestion that these alternative splices are functional. Experimental studies of many transcription factors have shown that deletion of the regulatory-interaction domain (leaving the DNA-binding domain intact) often results in a dominant phenotype, reversing the protein's normal effect on transcription even in the presence of the wild-type protein.<sup>24</sup>

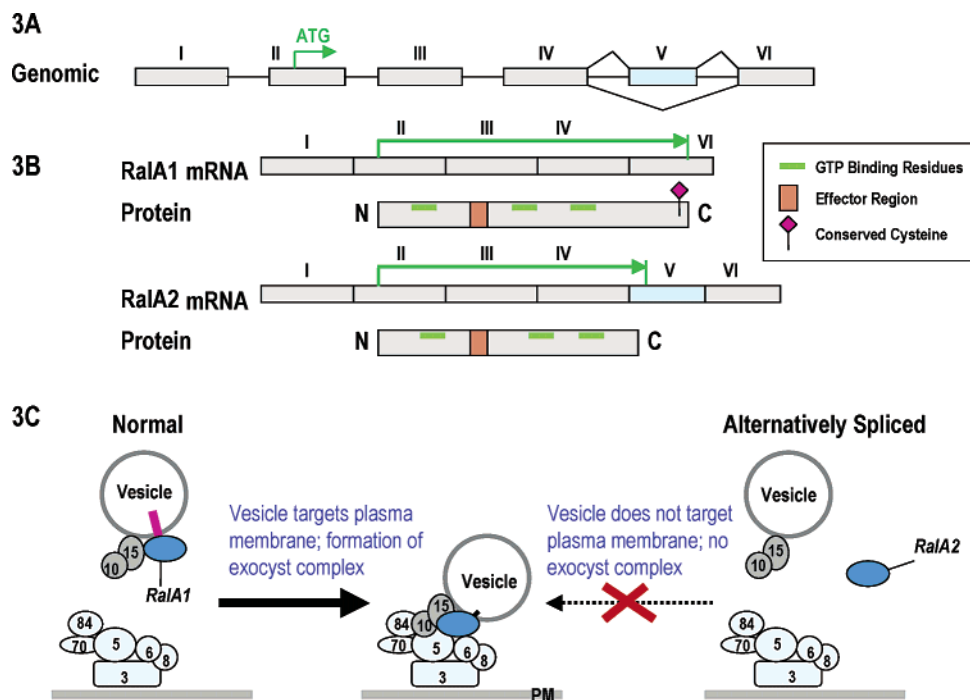
**Domain Splicing of the Transcriptional Regulator *Pbx2* Controls Protein Interactions with Transcriptional Co-Repressors.** The ASP database reveals a novel isoform of the transcriptional regulator *Pbx2*, in which alternative splicing removes its transcriptional repression domain (Figure 2A). Bioinformatics analysis identified Poly-Ala and homeobox domains present in both major and minor isoforms, as well as a protein-interaction domain that is removed in the novel isoform (Figure 2B). The protein-interaction domain is known to recruit the transcriptional corepressors NCoR and SMRT.<sup>25</sup> This domain disruption is similar to the pattern observed in the Kruppel family. Our analysis indicates that alternative splicing removes *Pbx2*'s repressive effect while retaining its target DNA sequence specificity. *Pbx* proteins are known to

form heterodimers with *Hox* family proteins, which in turn recruit transcriptional co-activators such as CBP (CREB binding protein).<sup>26,27</sup> It is possible that removal of the *Pbx2* C-terminal repression domain by alternative splicing activates transcription via the *Hox* protein.

This bioinformatics interpretation is strongly supported by independent experimental studies of the *Pbx* transcription factor family. An analogous pattern of alternative splicing of the *Pbx1* and *Pbx3* family members has been reported,<sup>28</sup> and in fact up-regulates target genes whose expression is normally repressed by full-length *Pbx* protein. Two functionally distinct protein isoforms of *Pbx1* (*Pbx1a* and *Pbx1b*) have been identified.<sup>25</sup> The long isoform *Pbx1a* possesses a C-terminal transcriptional repression domain, identical to that in *Pbx2*,<sup>25,29</sup> which has been shown to interact with the co-repressors SMRT and NCoR. Removal of the C-terminal *Pbx1a* repression domain switches off transcriptional repression and switches on transcriptional activation via the Hox-like protein *Pdb*, which recruits the co-activator CBP.<sup>25</sup>

Thus, changing protein-interaction networks by alternative splicing of protein interaction domains can redirect a biological pathway (Figure 2C). The discovery of the novel *Pbx2b* splice form, and the similar pattern in the Kruppel family, suggests alternative splicing plays an important role in transcriptional regulation via this specific pathway switching mechanism.

**Pathway Switching by an Abnormal Isoform of RalA GTPase.** We have identified a novel isoform of RalA, a Ras-



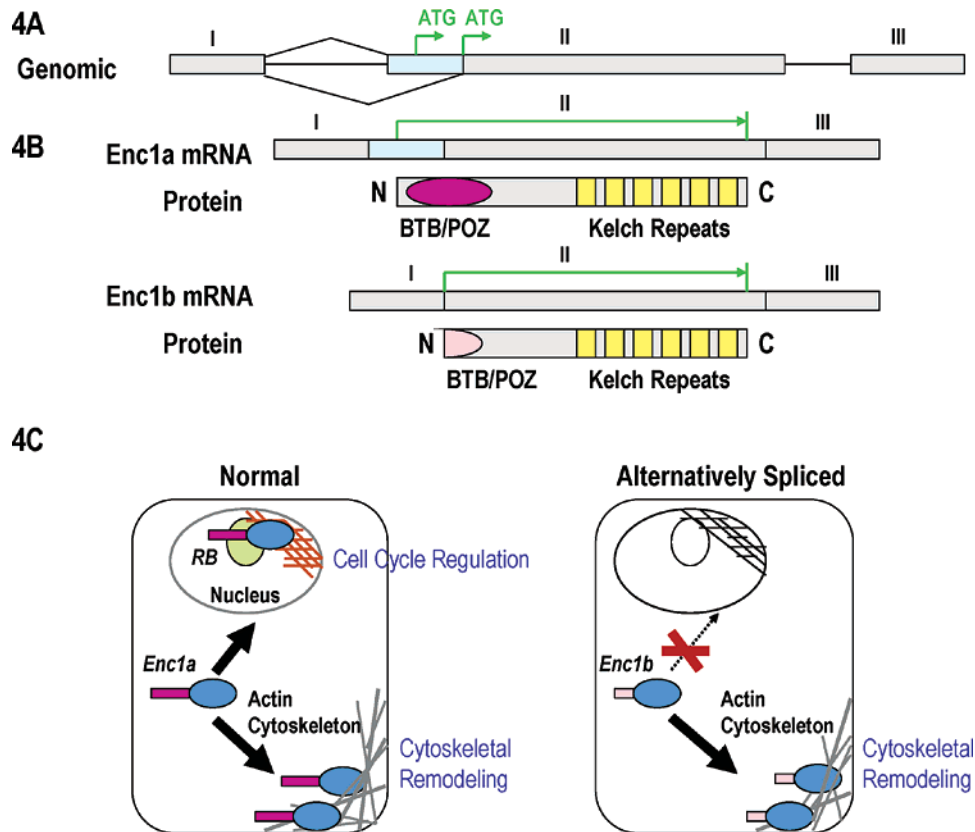
**Figure 3.** Alternative splicing of the RalA GTP-binding protein. (A) Genomic structure of the *RALA* gene. Exons are shown as gray boxes and the alternatively spliced exon is colored blue. The truncated isoform RalA2 is generated by an out-of-frame insertion of exonic material from exon V. (B) The two alternative forms of *RALA* inferred from the expressed sequence data. Alternative splicing removes a 4 amino acid “CCIL” conserved cysteine motif from the C-terminus of isoform RalA2. The protein-coding region for each isoform is represented by an arrow. (C) Proposed model of pathway switching by alternative splicing. Removal of the conserved RalA isoprenoid membrane attachment motif (magenta diamond in frame 1) provides a switching mechanism for the regulation of secretion and vesicle targeting. Membrane-bound RalA1 recruits the sec10 and sec15 subunits of the exocyst complex (adapted from<sup>51</sup>), facilitating vesicle targeting to the plasma membrane and permitting complete exocyst complex formation. Removal of the C-terminal isoprenyl attachment motif (frame 2) prevents vesicle targeting to the plasma membrane, thereby blocking formation of the complete exocyst complex.

related small GTPase involved in membrane trafficking and secretion in mammals.<sup>30</sup> Alternative splicing replaces the protein’s normal C-terminus (encoded by exon VI) by a novel C-terminus (encoded by a new exon V, which contains a STOP codon) (Figure 3A). This is supported by two independent ESTs from a colon-specific cDNA library. This removes a strongly conserved C-terminal CCIL sequence motif (Figure 3B), which is associated with covalent isoprenoid anchoring.<sup>31</sup> It is also possible that the premature STOP codon could induce nonsense-mediated decay (NMD), down-regulating the total amount of RalA transcript and protein in this tissue;<sup>32</sup> however, the repeated observation of this isoform as the dominant isoform in this tissue library suggests that it is still relatively abundant in this tissue. RalA has been shown to interact with the Sec5 subunit of the putative mammalian exocyst complex.<sup>30,33</sup> It appears that Sec5 is a direct target for activated RalA<sup>34,35</sup> and that this interaction is required for complete assembly of the exocyst complex. In yeast, the exocyst complex directs vectorial targeting of secretory vesicles to sites of rapid membrane expansion. In mammals, there is evidence that the exocyst complex directs Golgi-associated vesicles to the basolateral membrane of epithelial cells and to the growth cones of differentiating PC12 cells.<sup>35,36</sup> These data suggest that alternative splicing of RalA will interfere with its membrane localization and thus its interaction with the Sec5 exocyst complex and membrane trafficking pathway (Figure 3C). Indeed, RalA’s CCIL motif resembles the conserved membrane attachment motif found in p21, required for membrane localization of p21 via

posttranslational attachment of a 15-carbon farnesyl group.<sup>31,37</sup> RalA has been shown to undergo isoprenoid modification on the first conserved cysteine of the motif. The localization of RalA to the plasma membrane and to secretory or synaptic vesicle compartments is likely due to geranylation of its carboxy-terminus.<sup>38</sup> This modification promotes membrane attachment and is required for its involvement in membrane trafficking.<sup>30,33</sup>

**Domain Splicing of ENC1 Regulates Protein Interactions by Partial Removal of a Dimerization Domain.** We have also identified a novel splice variant of the ectodermal neuronal cortex-1 protein (ENC1), involved in neuronal differentiation in the developing and adult nervous system.<sup>39,40</sup> Bioinformatics analysis shows that alternative splicing removes a portion of the N-terminal BTB/POZ dimerization domain (Figure 4B) while leaving the rest of the protein unchanged, including 6 Kelch repeats. Although the precise effect of this change is unclear, our analysis suggests loss of dimerization and concomitant alteration of other protein interactions that depend on ENC1 dimerization. Several Kelch-containing proteins have been shown to bind actin,<sup>41</sup> and the BTB dimerization domain is thought to mediate protein–protein interactions associated with higher order structures involved in cytoskeletal remodeling and the regulation of nuclear processes such as chromatin remodeling and transcription.<sup>42</sup>

Alternative splicing of the BTB dimerization domain may direct localization of the ENC1 protein forms to different regions of the cell, altering protein–protein interactions im-



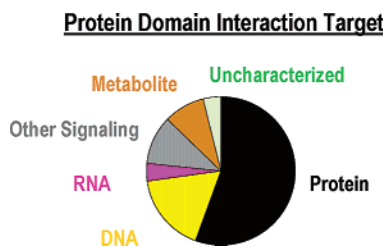
**Figure 4.** Alternative Splicing of the *Enc1* Cytoskeletal Protein. (A) Genomic structure of the *ENC1* gene. Exons are shown as gray boxes, and the alternatively spliced exonic region of exon II is colored blue. Alternative start codon usage results in the generation of 2 alternative isoforms, the shorter of which (*Enc1b*) is missing a portion of the BTB/POZ dimerization domain. (B) The two alternative forms of *ENC1* inferred from the expressed sequence data. The protein-coding region for each isoform is represented as a green arrow. (C) Proposed model of pathway switching by alternative splicing. Partial removal of the *Enc1* BTB/POZ dimerization domain may provide a switching mechanism for regulating the subcellular localization patterns, and subsequent cellular pathway roles of each isoform. The full-length *Enc1a* isoform (blue protein with dimerization domain colored magenta) can colocalize with the actin cytoskeleton (gray rodlike structures) in the cytoplasm, or bind the transcription factor  $RB^{p110}$  (green oval) in the nuclear matrix. The truncated *Enc1b* isoform (blue protein with partial dimerization domain colored pink) can bind to the actin cytoskeleton in the cytoplasm, but does not localize to the nucleus.

portant for cytoskeleton remodeling and cell cycle regulation (Figure 4C). A variety of experimental data support this hypothesis. First, Western blots of *ENC1* from primary neurons reveal 2 distinct protein bands of 67 and 57 kilodaltons,<sup>39</sup> matching the expected molecular weights of our major and minor protein isoforms (66 129 and 57 854 Daltons). Second, the two isoforms show different localization patterns in primary neurons. Although both forms were detected in cell lysates from primary neurons, only the full-length form was detected in the nuclear pellet. In the cytoplasm, full-length *ENC1* co-localizes with the actin cytoskeleton.<sup>39</sup> In the nucleus, full-length *ENC1* appears to localize to the nuclear matrix of the nucleoplasm, the peripheral heterochromatin and the nucleolus.<sup>39</sup> The nucleolus is known to harbor proteins involved in cell cycle regulation and differentiation,<sup>43</sup> and interestingly, *in vivo* and *in vitro* experiments indicate that full-length *ENC1* is involved with neuronal differentiation through its interaction with the cell cycle regulated, hypophosphorylated form of the retinoblastoma protein  $p110^{RB}$ .<sup>39,43</sup> These experiments suggest that this interaction is mediated by *ENC1*'s intact BTB dimerization domain. Our data predict that the short form of *ENC1* would lose this activity, while retaining its interaction with the actin cytoskeleton (via the Kelch repeats).

## Discussion

Our data suggest a pattern in which alternative splicing regulates biological function in a rather different way than transcriptional control. Transcriptional regulation has often been characterized as upregulating or downregulating a pathway, emphasizing a quantitative change (in total flux through a given pathway per unit time), while maintaining a static pathway structure (the basic network diagram remains unchanged). By contrast, alternative splicing appears to change the internal structure of a pathway by rearranging specific protein–protein interactions that control it. This hypothesis makes sense in terms of the distinct effects that these mechanisms can have on the proteome. Whereas transcriptional regulation changes the amount of a given protein, alternative splicing can actually alter the protein's internal structure, selectively modifying some of its interaction domains while retaining the others.

Consistent with this hypothesis, the major effect of alternative splicing in the human proteome appears to be addition or removal of protein–protein interaction domains (Figure 5). We manually classified alternatively spliced domain types according to the type of target molecule they bind: *protein–protein interaction domains*; *DNA-binding domains*; *RNA-binding do-*



**Figure 5.** Impact of alternative splicing on protein domain composition. Analysis of the types of protein domains removed by alternative splicing, classified by the type of molecule that each domain binds (see text).

mains; domains that bind *other signaling* ligands (such as cAMP, hormones, ligands and other signaling molecules); and domains that bind or act on a small-molecule substrate as part of metabolism or synthesis. The majority of alternatively spliced domains (56%) were protein–protein interaction domains. An identical analysis of alternatively spliced domains annotated in the SwissProt database produced a similar result (64% were protein–protein interaction domains (data not shown)). It should be emphasized that this simply reflects the ubiquitous occurrence of protein interaction domains in the human genome. Our analysis shows no statistically significant increase in the *rate* of alternative splicing of all protein–interaction domains relative to other types of domains, but does indicate that the majority of alternative splicing events alter protein interaction networks by adding or removing protein interaction domains. It should be noted that among specific domain classes, we have identified fifty kinds of domains that were alternatively spliced at a much higher rate (24–100%; see Table 1) than the rate observed for all protein domains on average (10.5%).

This effect may be important for understanding the regulation of functional pathways in organisms that have abundant alternative splicing. Recently, there has been great interest in protein interaction networks, both for elucidating their structure experimentally, characterizing them as networks, and using them to discover pathways.<sup>44–46</sup> It will be interesting to map the effects of alternative splicing onto the growing protein–interaction network databases, seeking to understand and predict alternative splicing’s effect on the proteome and on biological function. Unfortunately, at present most protein–interaction data is from organisms that have little alternative splicing (e.g., yeast and prokaryotes).

We have performed extensive validation tests on ASP,<sup>13</sup> which indicate that (1) the ASP isoform construction method works reliably from both mRNAs and ESTs, with an error rate for the construction method of under 2%. (2) Comparison of ASP and SwissProt sequences matched in 52 out of 57 genes tested (91%). (3) 78% of all ASP protein isoform sequences were validated by independent experimental literature, when we tested a sample of 20 genes. It is likely that some valid isoforms exist that have not yet been identified by published experimental data, so 100% validation is not expected at this time.

ASP offers a useful resource for biologists wondering how alternative splicing may affect their protein of interest. ASP provides a large, representative sample of alternative splicing’s proteomic impact ([www.bioinformatics.ucla.edu/ASP](http://www.bioinformatics.ucla.edu/ASP)). This could be useful for many studies, including identification of alternatively spliced protein forms by mass spectrometry,<sup>47–49</sup> and analysis of alternative splicing using protein–interaction network databases.<sup>44–46</sup> The sequence databases searched by

mass spectral software like RADARS<sup>48</sup> and ProSight PTM,<sup>49</sup> are capable of identifying a myriad of post-translational modifications including methylations, phosphorylations, lipoylations, and glycosylations to name a few. Such databases could benefit from the addition of protein isoform sequences, because these data can be used in combination with high-throughput mass spectrometry to assist in the identification of experimentally observed alternative splice forms (versus covalent modifications to a single protein form). The incorporation of ASP data into mass spectral studies could serve two primary functions: (1) provide experimental validation of putative protein forms and (2) enhance the existing pool of protein sequences searched by mass spectral software, thereby increasing the likelihood of obtaining a “hit” during the process of identifying peptide fragments.

ASP greatly expands available protein isoform data for the human proteome. Over half of ASP consists of novel isoforms not matching any mRNA or protein sequence deposited in GenBank. Currently, SwissProt<sup>50</sup> provides annotated alternative protein isoforms for 1989 human genes via its VARSPLIC feature. RefSeq currently has 932 reviewed human genes containing alternative transcript information,<sup>10</sup> although these data have a strong overlap with SwissProt. Adding the ASP data expands the dataset of alternatively spliced protein forms to a total of 5413 human genes (after removing overlaps between ASP and SwissProt), a nearly 3-fold increase. Finally, ASP is filtered to remove likely artifacts, and has been validated by a variety of tests. Thus, if a biologist finds an ASP isoform with interesting functional implications, our data indicate that experiments testing it would be worthwhile. The greatest benefit of such a database, in our view, would be to stimulate many specific experiments characterizing alternative splicing’s functional impact.

**Acknowledgment.** We wish to thank A. Courey, D. Eisenberg, C. Goulding, Q. Xu and T. Yeates for their helpful discussions and comments on this work. This work was supported by NIMH/NINDS Grant MH65166, NSF Grant No. 0082964, and DOE Grant No. DEFG0387ER60615. B.M. and R.R. were supported by NSF IGERT Award DGE-9987641.

**Supporting Information Available:** Conserved domain architectures of Kruppel family transcription factors (one Figure). This material is available free of charge via the Internet at <http://pubs.acs.org>.

## References

- (1) Mironov, A. A.; Fickett, J. W.; Gelfand, M. S. *Genome Res.* **1999**, *9*, 1288–1293.
- (2) Brett, D.; Hanke, J.; Lehmann, G.; Haase, S.; Delbruck, S.; Krueger, S.; Reich, J.; Bork, P. *FEBS Lett.* **2000**, *474*, 83–86.
- (3) Croft, L.; Schandorff, S.; Clark, F.; Burrage, K.; Arctander, P.; Mattick, J. S. *Nature Genet.* **2000**, *24*, 340–341.
- (4) Consortium., I. H. G. S. *Nature* **2001**, *409*, 860–921.
- (5) Kan, Z.; Rouchka, E. C.; Gish, W. R.; States, D. J. *Genome Res.* **2001**, *11*, 889–900.
- (6) Modrek, B.; Resch, A.; Grasso, C.; Lee, C. *Nucleic Acids Res.* **2001**, *29*, 2850–2859.
- (7) Xu, Q.; Modrek, B.; Lee, C. *Nucleic Acids Res.* **2002**, *30*, 3754–3766.
- (8) Schmucker, D.; Clemens, J. C.; Shu, H.; Worby, C. A.; Xiao, J.; Muda, M.; Dixon, J. E.; Zipursky, S. L. *Cell* **2000**, *101*, 671–684.
- (9) Kriventseva, E. V.; Koch, I.; Apweiler, R.; Vingron, M.; Bork, P.; Gelfand, M. S.; Sunyaev, S. *Trends Genet.* **2003**, *19*, 124–128.
- (10) Liu, S.; Altman, R. B. *Nucleic Acids Res.* **2003**, *31*, 4828–4835.
- (11) Lee, C.; Atanelov, L.; Modrek, B.; Xing, Y. *Nucleic Acids Res.* **2003**, *31*, 101–105.
- (12) Lee, C. *Bioinformatics* **2003**, *19*, 999–1008.

- (13) Xing, Y.; Resch, A.; Lee, C., submitted.
- (14) Schuler, G. *J. Mol. Med.* **1997**, *75*, 694–698.
- (15) Marchler-Bauer, A.; Panchenko, A. R.; Shoemaker, B. A.; Thiessen, P. A.; Geer, L. Y.; Bryant, S. H. *Nucleic Acids Res.* **2002**, *30*, 281–283.
- (16) Sonnhhammer, E. L.; Eddy, S. R.; Durbin, R. *Proteins* **1997**, *28*, 405–420.
- (17) Schultz, J.; Milpetz, F.; Bork, P.; Ponting, C. P. *Proc Natl Acad Sci USA* **1998**, *95*, 5857–5864.
- (18) Bayes, M.; Hartung, A. J.; Ezer, S.; Pispa, J.; Thesleff, I.; Srivastava, A. K.; Kere, J. *Hum. Mol. Genet.* **1998**, *7*, 1661–1669.
- (19) Miner, J. H. *Kidney Int.* **1999**, *56*, 2016–2024.
- (20) Chen, Y.; Sumiyoshi, H.; Oxford, J. T.; Yoshioka, H.; Ramirez, F.; Morris, N. P. *Matrix Biol.* **2001**, *20*, 589–599.
- (21) Collins, T.; Stone, J. R.; Williams, A. J. *Mol. Cell Biol.* **2001**, *21*, 3609–3615.
- (22) Barrera-Hernandez, G.; Cultraro, C. M.; Pianetti, S.; Segal, S. *Mol. Cell Biol.* **2000**, *20*, 4253–4264.
- (23) Berry, F. B.; Saleem, R. A.; Walter, M. A. *J. Biol. Chem.* **2002**, *277*, 10 292–10 297.
- (24) Klamt, B.; Koziell, A.; Poulat, F.; Wieacker, P.; Scambler, P.; Berta, P.; Gessler, M. *Hum. Mol. Genet.* **1998**, *7*, 709–714.
- (25) Asahara, H.; Dutta, S.; Kao, H. Y.; Evans, R. M.; Montminy, M. *Mol. Cell Biol.* **1999**, *19*, 8219–8225.
- (26) Chariot, A.; van Lint, C.; Chapelier, M.; Gielen, J.; Merville, M. P.; Bours, V. *Oncogene* **1999**, *18*, 4007–4014.
- (27) Chariot, A.; Princen, F.; Gielen, J.; Merville, M. P.; Franzoso, G.; Brown, K.; Siebenlist, U.; Bours, V. *J. Biol. Chem.* **1999**, *274*, 5318–5325.
- (28) Milech, N.; Kees, U. R.; Watt, P. M. *Genes Chromosomes Cancer* **2001**, *32*, 275–280.
- (29) van Dijk, M. A.; Peltenburg, L. T.; Murre, C. *Mech. Dev.* **1995**, *52*, 99–108.
- (30) Sugihara, K.; Asano, S.; Tanaka, K.; Iwamatsu, A.; Okawa, K.; Ohta, Y. *Nat. Cell Biol.* **2002**, *4*, 73–78.
- (31) Kinsella, B. T.; Erdman, R. A.; Maltese, W. A. *J. Biol. Chem.* **1991**, *266*, 9786–9794.
- (32) Lewis, B. P.; Green, R. E.; Brenner, S. E. *Proc. Natl. Acad. Sci. USA* **2003**, *100*, 189–192.
- (33) Brymora, A.; Valova, V. A.; Larsen, M. R.; Roufogalis, B. D.; Robinson, P. J. *J. Biol. Chem.* **2001**, *276*, 29 792–29 797.
- (34) Lipschutz, J. H.; Mostov, K. E. *Curr. Biol.* **2002**, *12*, R212–R214.
- (35) Moskalenko, S.; Henry, D. O.; Rosse, C.; Mirey, G.; Camonis, J. H.; White, M. A. *Nat. Cell Biol.* **2002**, *4*, 66–72.
- (36) Mostov, K. E.; Verges, M.; Altschuler, Y. *Curr. Opin. Cell Biol.* **2000**, *12*, 483–490.
- (37) Kinsella, B. T.; Maltese, W. A. *J. Biol. Chem.* **1991**, *266*, 8540–8544.
- (38) Feig, L. A.; Urano, T.; Cantor, S. *Trends Biochem. Sci.* **1996**, *21*, 438–441.
- (39) Kim, T. A.; Lim, J.; Ota, S.; Raja, S.; Rogers, R.; Rivnay, B.; Avraham, H.; Avraham, S. *J. Cell Biol.* **1998**, *141*, 553–566.
- (40) Kim, T. A.; Ota, S.; Jiang, S.; Pasztor, L. M.; White, R. A.; Avraham, S. *Gene* **2000**, *255*, 105–116.
- (41) Adams, J.; Kelso, R.; Cooley, L. *Trends Cell Biol.* **2000**, *10*, 17–24.
- (42) Rando, O. J.; Zhao, K.; Crabtree, G. R. *Trends Cell Biol.* **2000**, *10*, 92–97.
- (43) Ferguson, K. L.; Slack, R. S. *Neuroreport* **2001**, *12*, A55–A62.
- (44) Marcotte, E. M.; Pellegrini, M.; Thompson, M. J.; Yeates, T. O.; Eisenberg, D. *Nature* **1999**, *402*, 83–86.
- (45) Duan, X. J.; Xenarios, I.; Eisenberg, D. *Mol. Cell. Proteomics* **2002**, *1*, 104–116.
- (46) Aloy, P.; Russell, R. B. *Proc. Natl. Acad. Sci. USA* **2002**, *99*, 5896–5901.
- (47) Shevchenko, A.; Sunyaev, S.; Loboda, A.; Bork, P.; Ens, W.; Standing, K. G. *Anal. Chem.* **2001**, *73*, 1917–1926.
- (48) Field, H. I.; Fenyó, D.; Beavis, R. C. *Proteomics* **2002**, *2*, 36–47.
- (49) Taylor, G. K.; Kim, Y. B.; Forbes, A. J.; Meng, F.; McCarthy, R.; Kelleher, N. L. *Anal. Chem.* **2003**, *75*, 4081–4086.
- (50) Bairoch, A.; Apweiler, R. *Nucleic Acids Res.* **1998**, *26*, 38–42.
- (51) Guo, W.; Sacher, M.; Barrowman, J.; Ferro-Novick, S.; Novick, P. *Trends Cell Biol.* **2000**, *10*, 251–255.

PR034064V