blood®
Regular Article

RED CELLS, IRON, AND ERYTHROPOIESIS

A cytosine-rich splice regulatory determinant enforces functional processing of the human α-*globin* gene transcript

Xinjun Ji,[1] Jesse Humenik,[1] and Stephen A. Liebhaber[1,2]

[1]Department of Genetics and [2]Department of Medicine, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA

KEY POINTS

- **A C-rich determinant in intron 1 enforces functional splicing of the *hα-globin* transcript.**
- **The splice regulatory function of the C-rich determinant is achieved through interactions with polyC-binding proteins.**

The establishment of efficient and stable splicing patterns in terminally differentiated cells is critical to maintenance of specific functions throughout the lifespan of an organism. The human *α-globin* (*hα-globin*) gene contains 3 exons separated by 2 short introns. Naturally occurring α-thalassemia mutations that trigger aberrant splicing have revealed the presence of cryptic splice sites within the *hα-globin* gene transcript. How cognate (functional) splice sites are selectively used in lieu of these cryptic sites has remained unexplored. Here we demonstrate that the preferential selection of a cognate splice donor essential to functional splicing of the *hα-globin* transcript is dependent on the actions of an intronic cytosine (C)-rich splice regulatory determinant and its interacting polyC-binding proteins. Inactivation of this determinant by mutation of the C-rich element or by depletion of polyC-binding proteins triggers a dramatic shift in splice donor activity to an upstream, out-of-frame, cryptic donor. The essential role of the C-rich element in *hα-globin* gene expression is supported by its coevolution with the cryptic donor site in primate species. These data lead us to conclude that an intronic C-rich determinant enforces functional splicing of the *hα-globin* transcript, thus acting as an obligate determinant of *hα-globin* gene expression. (*Blood*. 2019;133(21):2338-2347)

Introduction

Posttranscriptional controls play a major role in the regulation of eukaryotic gene expression.[1] These controls are mediated by specific interactions of *cis*-acting sequences and structures on target transcripts and/or their processed messenger RNAs (mRNAs) with *trans*-acting RNA-binding proteins and/or non-coding RNAs.[2] RNA splicing controls comprise a major subset of posttranscriptional gene regulatory determinants.[3] Interruption of normal splicing patterns can significantly affect gene expression. In the case of the globin genes, mutations that interfere with normal splicing result in loss of globin protein expression and a corresponding set of α-thalassemia syndromes.[4] The array of structural determinants in the *hα-globin* transcript that ensures the generation of functional *hα-globin* mRNAs remains to be more fully defined.

The critical importance of RNA-protein interactions to the expression of the *hα-globin* gene has been demonstrated by studies of the polycytosine (C)-binding proteins (PCBPs) PCBP1 and PCBP2. These proteins comprise a subset of KH domain RNA-binding proteins with high-specificity and high-avidity C-rich, pyrimidine-pure motifs.[5,6] We have previously demonstrated that PCBPs can affect the nuclear processing as well as cytoplasmic stability of the *hα-globin* gene transcript.[7-10] These controls are mediated via binding to an array of C-rich elements located within the nascent transcript and in the mature *hα-globin* mRNA.[8-13] For example, binding of PCBPs to the C-rich site within the 3′ UTR of the *hα-globin* mRNA modulates splicing and 3′ cleavage/polyadenylation of the nascent *hα-globin* transcript in the nucleus[9,10] and enhances the stability of *hα-globin* mRNA[7,8,11,12] in the cytoplasm.

The *hα-globin* transcript is normally processed via a constitutive splicing pathway in which the 2 short introns are neatly excised and the 3 exons are spliced together to generate a functional and efficiently translated mRNA. The accuracy of this splicing pathway is essential to the high-level expression of hα-globin protein. In contrast to a large majority of mammalian transcripts, alternative splicing does not seem to be involved in globin gene expression. Although extensive studies have focused on the roles of RNA-binding protein interactions with target transcripts in the modulation of alternative splicing pathways, much less emphasis has been placed on understanding how constitutive splicing patterns, such as those involved in globin gene expression, are established and enforced. Importantly, consensus sequences for splice sites are highly degenerate and are often present at multiple sites that are not used to a significant extent (ie, cryptic splice sites).[14-16] Therefore, understanding how

cognate sites are selected over cryptic sites is critical to a full understanding of eukaryotic gene regulation in general and globin gene expression in particular.

The thalassemia syndromes arise from a large and complex set of defects in globin gene expression.[4] A defined subset of thalassemia mutations affect *hα-globin* transcript splicing.[17-22] Analysis of these splicing mutations was among the first to facilitate mapping of the *cis*-acting sequence determinants of the splicing pathway and the first to reveal that gene mutations can activate cryptic sites within a transcript.[23] What has remained relatively unexplored is how functional sites in the globin transcripts are selectively used to the exclusion of cryptic splice sites to support effective high-level globin gene expression.

In the current report, we explore the basis for constitutive splicing of the *hα-globin* gene transcript. These studies reveal that utilization of the functional (cognate) intron 1 splice donor site is dependent on the function of a closely positioned intronic C-rich splice regulatory determinant and its interaction with 1 or more polyC-binding proteins.

Methods

Cell culture and transfection

Human erythroleukemia (K562) cells were grown in RPMI 1640 medium. MEL cells and HeLa cells with a stably expressed transfected tet-off transactivator (MEL/tTA, HeLa/tTA) were used for conditional expression of the *hα-globin* mRNA.[7] All culture media contained 100 U/mL of penicillin and 100 μg/mL of streptomycin sulfate, and conditions were maintained at 37°C in a 5% carbon dioxide incubator. MEL/tTA cells were transfected with each of an indicated pTet plasmid DNA by electroporation.[7,24] The HeLa/tTA cell transfections were carried out using the liposomal reagent Trans-IT (Mirus).[7] Cells were then cultured in tet$^-$ media for 24 hours to induce expression from the transfected pTet plasmid. K562 cells were transfected using Nucleofector V (Amaxa) as previous described.[25]

Sucrose gradients

Ten percent to 50% and 10% to 40% linear sucrose gradient fractionations were performed as described.[24]

IP

K562 cells were washed with ice-cold phosphate-buffered saline twice and resuspended in 1000 μL of ice-cold RSB100 buffer (10 mM of Tris hydrogen chloride [pH, 7.4], 100 mM of sodium chloride, and 2.5 mM of magnesium chloride) containing 0.5% Triton-X-100 on ice for 5 minutes. Cytoplasmic cell extracts were prepared by centrifugation[24] and used for immunoprecipitation (IP) experiments. IP was carried out with affinity-purified antibody to PCBP1 and PCBP2 or with preimmune serum, all as described previously.[24] IP pellets were extracted and ethanol precipitated before RNase protection assay (RPA) or reverse transcription polymerase chain reaction (RT-PCR) analysis. Primary antibodies to the PCBP isoforms have been previously characterized.[26]

UV crosslinking assay and IP assay

Wild-type (WT) and mutated RNA oligonucleotides were 5′-end labeled using T4 polynucleotide kinase (NEB, Beverly, MA) and [γ-^{32}P]ATP (Amersham). The labeled oligonucleotides were gel purified before use; 5 ng of each oligonucleotide (~20 000 cpm) was incubated with HeLa cell nuclear extract in a 25-μL reaction containing 60% of nuclear extract (15 μL) and 1 mM of EDTA at 30°C for 20 minutes. The reactions were subsequently irradiated for 10 minutes at 254 nm at 4°C. IPs were carried out as described in the previous paragraph. After the addition of sodium dodecyl sulfate–loading buffer, the samples were analyzed by sodium dodecyl sulfate polyacrylamide gel electrophoresis.

In vitro splicing assays

In vitro splicing was performed as described.[9] After assay incubation, the reaction was phenol-chloroform extracted and ethanol precipitated, and the pellet was resuspended in loading buffer (Figure 3) or diethyl pyrocarbonate–treated water (Figure 4) for the RT-PCR assay. In vitro splicing assays using the polyC-depleted nuclear extract (Figure 6) were performed as previously described.[9]

RPA

Internally labeled ^{32}P-probes used for RPA were generated by in vitro transcription of plasmids containing partial inserts for *hα-globin*[24] and *hGAPDH* (Ambion, Austin, TX) using a Maxiscript SP6 kit under conditions recommended by the manufacturer (Ambion). RPA was carried out as described previously.[24]

RT-PCR

RT-PCRs were performed as described.[9,27] Briefly, Trizol-extracted total RNAs were treated with DNase I (amplification grade; Invitrogen) and then reverse transcribed using oligo-dT, Moloney murine leukemia virus reverse transcriptase (Promega), and 1× Moloney murine leukemia virus RT buffer (Promega) according to manufacturers' instructions. After incubation at 37°C for 1 hour, the samples were used as a template for PCR. Primers used are as follows: hα-globin forward, 5′-ACTCTTC TGGTCCCCACAGACTCA-3′; hα-globin reverse, 5′-CAGGGC GTCGGCCACCTTCTTG-3′.

Minigene analysis

WT and mutant minigenes (Figure 5) were cloned into the PCDNA3 vector between *Eco*R I and *Xho*I cloning sites. Each minigene contains 2 parts; the upstream segment contains hα-globin exon 1 (132 bp) and extends for 80 bp into the contiguous intron 1, and the second segment contains partial intron (43 bp) and the full downstream exon (106 bp) of the pI-11(-H3)-PL splicing minigene plasmid.[28] Transfections of minigene plasmids into K562 cells were performed as described. RT-PCR was performed with SP6 and T7 primers, using neomycin mRNA expressed from PCDNA3 vector as internal control. Primers for neomycin mRNA are as follows: neomycin-F, 5′-TTGTCACTG AAGCGGGAAGG-3′; neomycin-R, 5′-ATGCGATGTTTCGCTT GGTG-3′.

Statistics

Statistical significance (*P* values) was determined using 2-tailed, unpaired Student *t* test.

Results

Detection of a cryptic splice donor site within exon 1 of the *hα-globin* transcript

In the course of our studies of globin mRNA translational controls, we carried out a sucrose gradient fractionation of polysomes

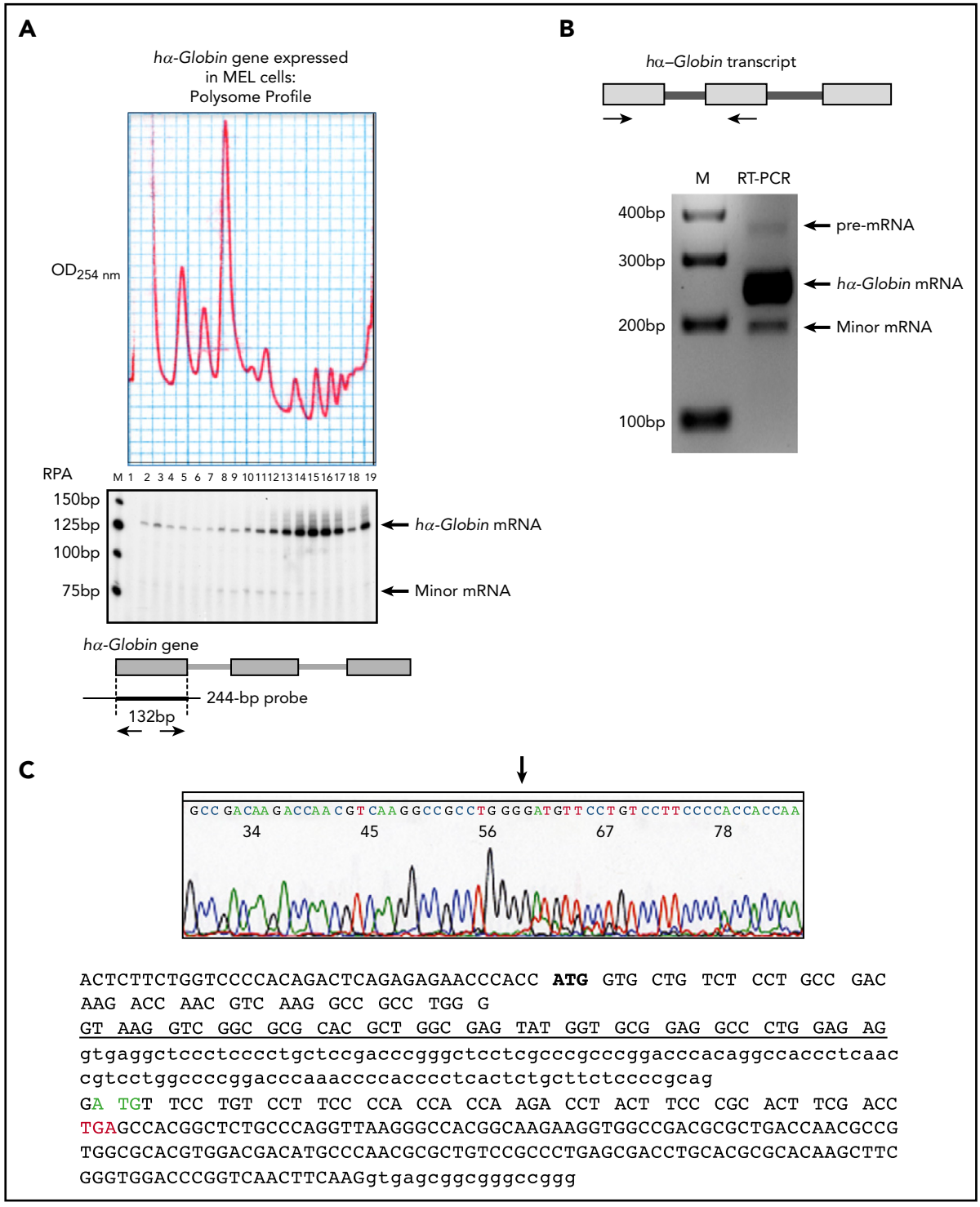

Figure 1. Detection and structural definition of a minor splicing product from the *hα-globin* gene. (A) MEL/tTA cells transfected with the *hα-globin* gene express a predominant full-length *hα-globin* mRNA and a minor species with distinct polysome profiles. MEL/tTA cells[7] were transfected with a plasmid p(Tet-αWT) expressing the full-length *hα-globin* gene.[7] RPA was performed across the sucrose gradient. A diagram of the RPA is shown below the gel image; the entire 3-exon α-globin gene is shown along with its 2 introns. The upper band (132 bp) in the gel (*hα-globin* mRNA) represents protection of the ^{32}P internally labeled 244-bp probe by the *hα-globin* transcript spliced at the cognate exon 1/intron 1 splice donor site. The smaller protected fragment of ~80 bp (minor mRNA) suggests the presence of a small population of *hα-globin* RNA generated by an alternative processing pathway. The distribution of the minor RNA is shifted strongly to the left (ie, to lighter fractions) compared with the full-length *hα-globin* mRNA. A 25-bp size marker ladder (M) is shown on the left of the gel; 19 fractions are labeled on the top of the gel and align with the sucrose gradient tracing (OD254 nm). (B) Amplification of *hα-globin* complementary DNAs from a transfected *hα-globin* gene in MEL/tTA cells. RNA was isolated from cells 48 hours posttransfection. Three bands were generated by RT-PCR with a primer set bracketing intron 1 (horizontal arrows; top diagram); the largest band (355 bp) corresponds to the unprocessed *hα-globin* transcript with intron 1 (pre-mRNA), the second band (238 bp) corresponds to the *hα-globin* mRNA subsequent to excision of intron 1 from the cognate donor site (*hα-globin* mRNA), and the third band (189 bp; minor mRNA) was of unknown structure. The minor DNA fragment was excised and sequenced (bottom). M indicates 100-bp size marker ladder. (C) The *hα-globin* transcript undergoes low-frequency splicing from a cryptic splice donor site located within exon 1. The sequence of the minor mRNA revealed utilization of a cryptic splice donor located within exon 1 ligated to the cognate splice acceptor of exon 2. Sequences are shown at the bottom of the figure; capital letters indicate exons 1 and 2, and lower case

isolated from MEL/tTA cells transfected with an inducible *hα-globin* gene[7] (Figure 1). We profiled *hα-globin* mRNA across the gradient with a 244-nt RPA probe that encompassed exon 1 and extended 5′ into the promoter region and 3′ into the adjacent intron 1 (Figure 1A bottom). This analysis revealed a predominant protected fragment of 132 nt that corresponded in size to the properly spliced exon 1. This band distributed across the actively translating polysomes in a pattern similar to that previously defined for the mature *hα-globin* mRNA.[24] Unexpectedly, we also observed trace levels of a second fragment of 83 nt (ie, minor mRNA). This minor RNA species was restricted to the monosome and disome regions of the gradient. A parallel polysome analysis of the endogenous *hα-globin* mRNA in K562 cells revealed the same 2 RNAs with the same relative representations and polysome distributions as noted in the MEL cell study (supplemental Figure 1A, available on the *Blood* Web site). RNAs corresponding to the 2 protected fragments were both present in ribonucleoprotein (RNP) complexes immunoprecipitated from K562 cells with an antibody to PCBP2 (supplemental Figure 1B). Because PCBP2 binding to the mature *hα-globin* mRNA is limited to a unique C-rich motif within the 3′ UTR,[7-9,24,26] these data suggested that the RNA represented by the short RPA product extended into the *hα-globin* 3′ UTR (supplemental Figure 1B). These data led us to conclude that the *hα-globin* transcript is subject to a low-efficiency minor splicing pathway, generating trace levels of a *hα-globin* mRNA containing a deletion within exon 1.

The exact structure of the minor *hα-globin* mRNA, determined by RT-PCR amplification and sequencing (Figure 1B-C), revealed that it was generated by low-efficiency splicing between a cryptic splice donor located within exon 1 and the canonical exon 2 splice acceptor (Figure 1C). This cryptic splice donor within exon 1 is the same site as that activated secondary to a naturally occurring 5-bp α-thalassemia deletion that removes the canonical intron 1 donor site.[17,29] Thus, the exon 1 cryptic site is shown to be weakly active in the WT *hα-globin* mRNA.

C-rich element within intron 1 drives the preferential use of the cognate splice donor

The basis for the predominant utilization of the major exon 1 splice donor as compared with the cryptic splice donor could not be readily explained on the basis of differences in their primary sequences, both of which were well aligned to the splice donor consensus as assessed by 2 independent algorithms (supplemental Figure 2).[17,29]

We have previously reported that the PCBP (also referred to as αCP or hnRNP E) binds to an extensive C-rich tract that encompasses the lariat branch point at the intron 1 splice acceptor of the *hα-globin* transcript and represses splicing.[9] In that study, we also identified, but did not further characterize, a distinct C-rich tract located immediately 3′ to exon 1 splice donor within intron 1.[9] The proximity of this C-rich segment to the splice donor suggested that it might affect intron 1 splice site selection/utilization. To test this model, we first confirmed that PCBPs could bind in a sequence-specific manner to this C-rich segment by incubating ^{32}P-labeled oligonucleotides corresponding to the C-rich region or to the corresponding region with 2 C→G substitutions with HeLa cell extracts (Figure 2A). These incubations were UV crosslinked and immunoprecipitated with antibodies to PCBP1, PCBP2, and preimmune immunoglobulin G (Figure 2B). The WT probe assembled an RNP complex that contained PCBP1 and PCBP2, as evidenced by the IPs with the respective isoform-specific antibodies (Figure 2B). The C→G substitutions within the C-rich tract blocked PCBP RNP complex formation. These in vitro binding studies confirmed that the C-rich motif adjacent to the cognate intron 1 splice donor can be targeted by 1 or more polyC-binding proteins.

Given the prominent roles of RNA-binding proteins in splicing regulation, we next hypothesized that assembly of a polyC-RNP complex adjacent to the cognate intron 1 donor site might contribute to the predominant use of the cognate vs cryptic splice donor. This model was initially tested in an in vitro splicing assay (Figure 3). An *hα-globin* RNA splicing substrate was generated that extended from exon 1 through intron 1 and into exon 2 (Figure 3A). When this WT probe was incubated in a nuclear extract optimized for in vitro splicing activity, we observed the generation of the normally spliced product at a level of 24% input. Remarkably, a parallel reaction on an RNA substrate covering the same C-rich region but containing the 2 C→G substitutions (mutation 1) substantially reduced cognate donor utilization (24% to 15% input) and reciprocally activated the cryptic splice site (from trace to 4% of input; Figure 3B left panel). Of note, the 2-base C→G substitutions did not alter the optimal sequence configuration of the splice donor itself (supplemental Figure 3). These data supported the model in which this C-rich segment enforces use of the cognate intron 1 donor site.

We have previously reported that mutations of a C-rich region overlying the branch point site of the intron 1 splice acceptor (mutation 3) have a repressive impact on the activity of the intron 1 splicing. To test the independent functioning of the donor site C-rich element, we assessed any impact that the C-rich branch point region might have on splice donor selection. Consistent with our prior studies,[9] we observed that mutation of the C-rich region encompassing the intron 1 branch point (mutation 3) enhanced (derepressed) overall intron splicing activity (compare mutations 13 and 123 with mutations 1 and 12, respectively). Importantly, however, the branch point (mutation 3) substitutions had no appreciable impact on the relative utilization of the competing cognate and cryptic splice donors (Figure 3B right panel). In contrast, and consistent with our previous report,[9] a set of C→T substitutions located 3′ to the lariat branch point (mutation 2) had no impact on splicing activity (compare mutations 1 and 12). These data led us to conclude that the C-rich motif adjacent to the cognate intron 1 splice donor enforces the utilization of the cognate site and that this splice control activity is independent of the distinct C-rich region encompassing the intron 1 acceptor site.

Figure 1 (continued) letters indicate sequences in introns 1 and 2. The region of exon 1 that is converted to intronic sequence by the use of the cryptic splice donor within exon 1 is underlined. The splicing from this cryptic splice donor generates a shorter *α-globin* mRNA that would be out of frame with the WT *α-globin* mRNA with a new in-frame stop codon within exon 2 (TGA; red font). This shorter *α-globin* mRNA also contains a second potential translation initiation site within exon 2 (green font) that would be in frame with the WT *hα-globin* mRNA.

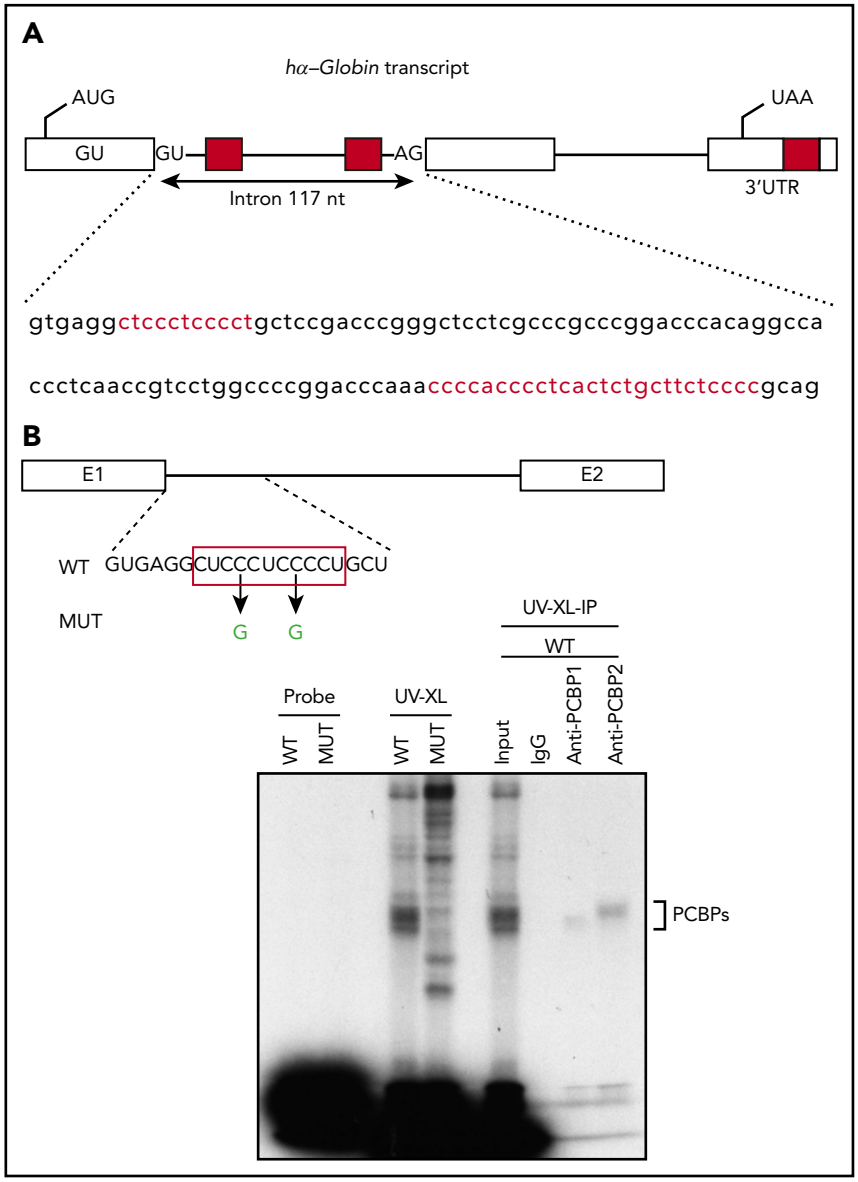

Figure 2. The *hα-globin* transcript contains a C-rich polypyrimidine tract located immediately 3′ to the cognate intron 1 splice donor that is a binding target for PCBP1 and PCBP2). (A) The *hα-globin* transcript and an expanded view of the intron 1 sequence. The *hα-globin* gene transcript contains 3 exons; the start codon in exon 1 and the stop codon in exon 3 are indicated. There are 3 defined C-rich tracts within the *hα-globin* transcript, each indicated by a red box: a C-rich element immediately 3′ to the cognate splice donor of intron 1, a second intronic C-rich element overlying the branch point polypyrimidine tract of the splice acceptor of intron 1, and a third C-rich tract is located within the 3′ UTR. The full intron 1 sequence is shown below the gene diagram, with the 2 C-rich tracts highlighted in red font. The C-rich region at the polypyrimidine tract of intron 1 and that within the 3′ UTR have both been previously linked to functions in transcript processing[10,25,27] and mRNA stability.[7,8,11,12] The function of the C-rich tract located immediately 3′ to the cognate intron 1 splice donor site has not been previously assessed for function. (B) The C-rich tract 3′ to the cognate intron 1 splice donor site of the *hα-globin* transcript is recognized by PCBP1 and PCBP2. A diagram of the region of the *hα-globin* transcript extending from exon 1 into exon 2 is shown, and the C-rich segment immediately 3′ to the cognate splice donor is boxed in red. Two Cs within this segment were converted to Gs to test for C-dependent activity of this region. The WT and mutant (MUT) probes were labeled with ^{32}P and incubated with HeLa cell extracts. After UV crosslinking, the extracts were analyzed on a sodium dodecyl sulfate polyacrylamide gel electrophoresis gel. A prominent complex (PCBPs) formed on the WT but not on the MUT probe; this complex comigrates with the complexes in the crosslinked extract that were immunoprecipitated with affinity-purified isotype-specific antiseras to PCBP1 and PCBP2. IgG, immunoglobulin G.

We have demonstrated in prior reports that a C-rich region in the 3′ UTR can exert a long-range impact on the activity of intron 1 splicing.[9] With that in mind, we next asked whether this 3′ UTR C-rich element played a role in cognate vs cryptic splice donor activity. To this end, we compared in vitro splicing patterns of the full-length *hα-globin* transcript with and without the 3′ UTR C-rich element (Figure 4) (α^{WT} and α^{Neut}).[7,8] The result of these studies served 2 purposes; they confirmed in the context of the full-length *hα-globin* transcript that the C-rich motif adjacent to the cognate intron 1 splice donor enforces utilization of the cognate splice donor site, and they further demonstrated that the splicing control activity of this determinant is independent of the C-rich region in the 3′ UTR (Figure 4). These finding were fully validated in a separate study in which intact cells were transfected with plasmids encoding full-length *hα-globin* mRNA with the full set of mutations described for the in vitro splicing studies (supplemental Figure 4). Taken together, these in vitro and cell-based studies demonstrate the essential role of the C-rich region 3′ to the intron 1 splice donor in enforcing a functional splicing pattern of the *hα-globin* gene transcript and further demonstrate that this function is independent of the C-rich motifs at the intron 1 splice acceptor site and within the 3′ UTR.

C-rich splice regulatory determinant establishes the proper splicing pattern of *hα-globin* transcript via its impact on donor site competition

Having established that the C-rich sequence at the intron 1 donor site is important to maintain the proper expression of the *hα-globin* gene, we next ask how this is achieved. To specifically focus on the intron 1 donor activity, we prepared a series of minigene constructs (Figure 5A-B) that isolated the competing splice donors from the rest of the *hα-globin* gene. These constructs were designed to selectively disrupt 3 regions either individually or in various combinations: the C-rich motif adjacent to the intron 1 donor site (Figure 3), the cognate intron 1 splice donor site itself (introduction of a naturally occurring α-thalassemia mutation[17,29]), and the competing cryptic donor within exon 1.

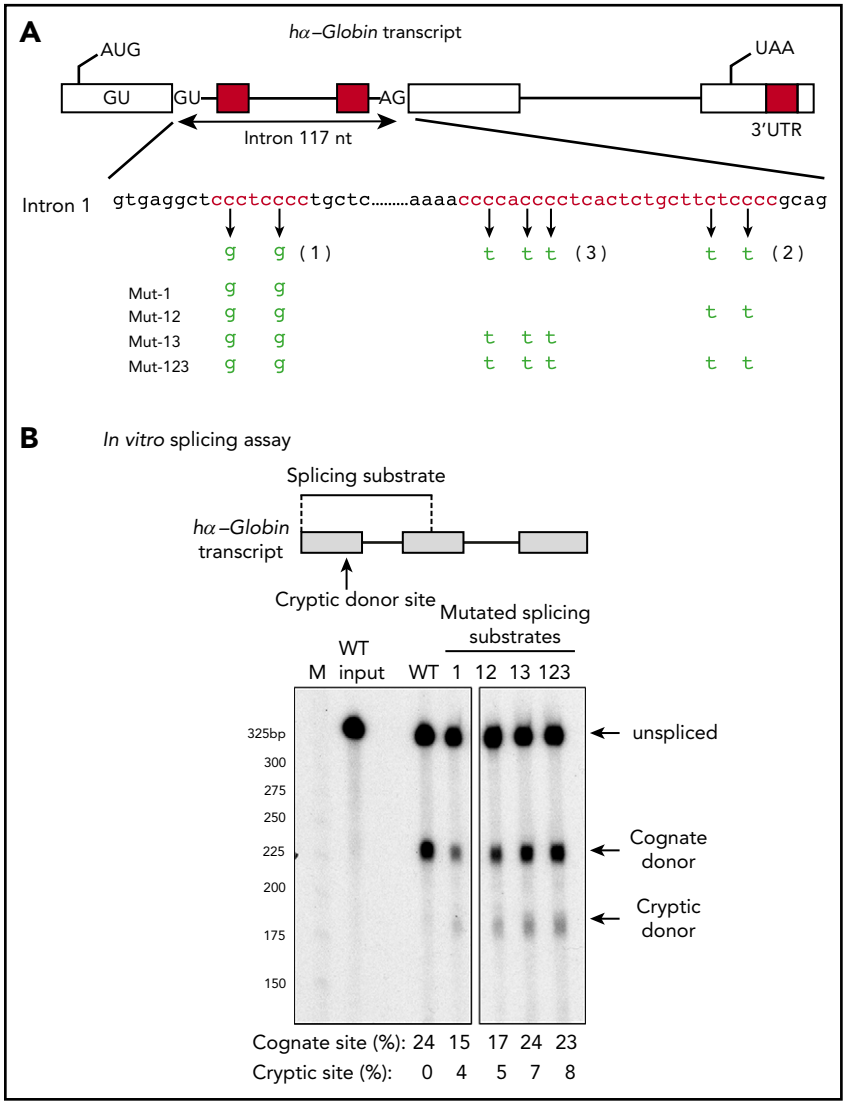

Figure 3. The C-rich motif 3′ to the intron 1 splice donor enforces functional splicing of the *hα-globin* transcript. (A) Mutations introduced in intron 1. The *hα-globin* gene diagram and the C-rich segments are as described in Figure 2A. Mutations that interrupt the C-rich elements within intron 1 are indicated in green font. The combinations of mutations listed below the sequence were assayed for impact on splicing in vitro. (B) In vitro splicing analysis reveals that the C-rich tract adjacent to the cognate splice donor enforces functional splicing of exon 1. Each ^{32}P internally labeled RNA substrate was incubated with HeLa nuclear extract. The products of each in vitro splicing reaction were assessed by denaturing polyacrylamide gel electrophoresis. The results reveal that mutations in the C-rich tract adjacent to the cognate splice donor (mutation 1 [Mut 1]) accentuates splicing from the cryptic site with a reciprocal decrease in normally spliced RNA. Additional mutations within the C-rich tracts adjacent to the splice acceptor fail to alter the activation of the cryptic donor by Mut 1 (also described in "Results"). The differential usages of the 2 donor sites (shown below the gel) are calculated as percentages of the individual band intensities (cognate or cryptic site splicing) over the total RNA (unspliced plus cognate site splicing plus cryptic site splicing). WT input substrate alone (not incubated with HeLa cell extract) is shown. M indicates 25-bp size marker ladder.

Each minigene was expressed in K562 cells, and the corresponding transcripts were assayed for splice donor utilization. The analyses revealed that a majority of the mRNAs generated from the WT transcript were spliced from the cognate donor site, with only a minor portion of mRNAs originating from usage of cryptic splicing donor (WT; Figure 5C). This pattern fully reproduced that observed for the native *hα-globin* transcript in K562 cells (supplemental Figure 1). Selective disruption of the C-rich site adjacent to the cognate donor (2 C→G replacements) resulted in a dramatic shift to utilization of the cryptic donor (C→G; Figure 5C). When the cognate donor site was selectively inactivated, the splicing was fully shifted to the cryptic donor site (ΔTGAGG alone or ΔTGAGG + C→G; Figure 5C), as occurs when this same mutation is present in an individual with α-thalassemia.[17,29] Direct inactivation of the cryptic donor site generated mRNAs exclusively from the cognate splicing donor site (ΔTAA alone or ΔTAA + C→G; Figure 5C). Comparison of transcripts that contained only the cryptic (ΔTGAGG) or only the cognate splice donor site (ΔTAA) demonstrated that the 2 donors were of comparable strength when not in direct competition. This equivalency of splice donor strength is consistent with their equivalent sequence match to optimal splice donor motifs (supplemental Figures 2 and 3). These minigene assays support the model that the predominant use of the cognate donor site in the *hα-globin* transcript is dependent on the actions of the adjacent C-rich splice regulatory determinant. Although this element seems to have a small direct enhancing impact on the cognate donor when this site is studied in isolation, a far more dramatic impact on splice site selection is observed when the cognate splice donor is competing with the cryptic site.

Actions of the splice regulatory determinant are achieved through the actions of 1 or more polyC-binding proteins

Because the predominant use of the cognate exon 1 splice donor sites is strongly affected by the adjacent C-rich determinant (Figures 3 and 4), and this determinant can be bound by 2 defined PCBPs (Figure 2), we next asked whether the observed splice regulatory activity of the determinant was dependent on the actions of PCBPs. This was tested by in vitro splicing analysis of the full-length *hα-globin* transcript in HeLa nuclear extracts selectively depleted of PCBPs (Figure 6).

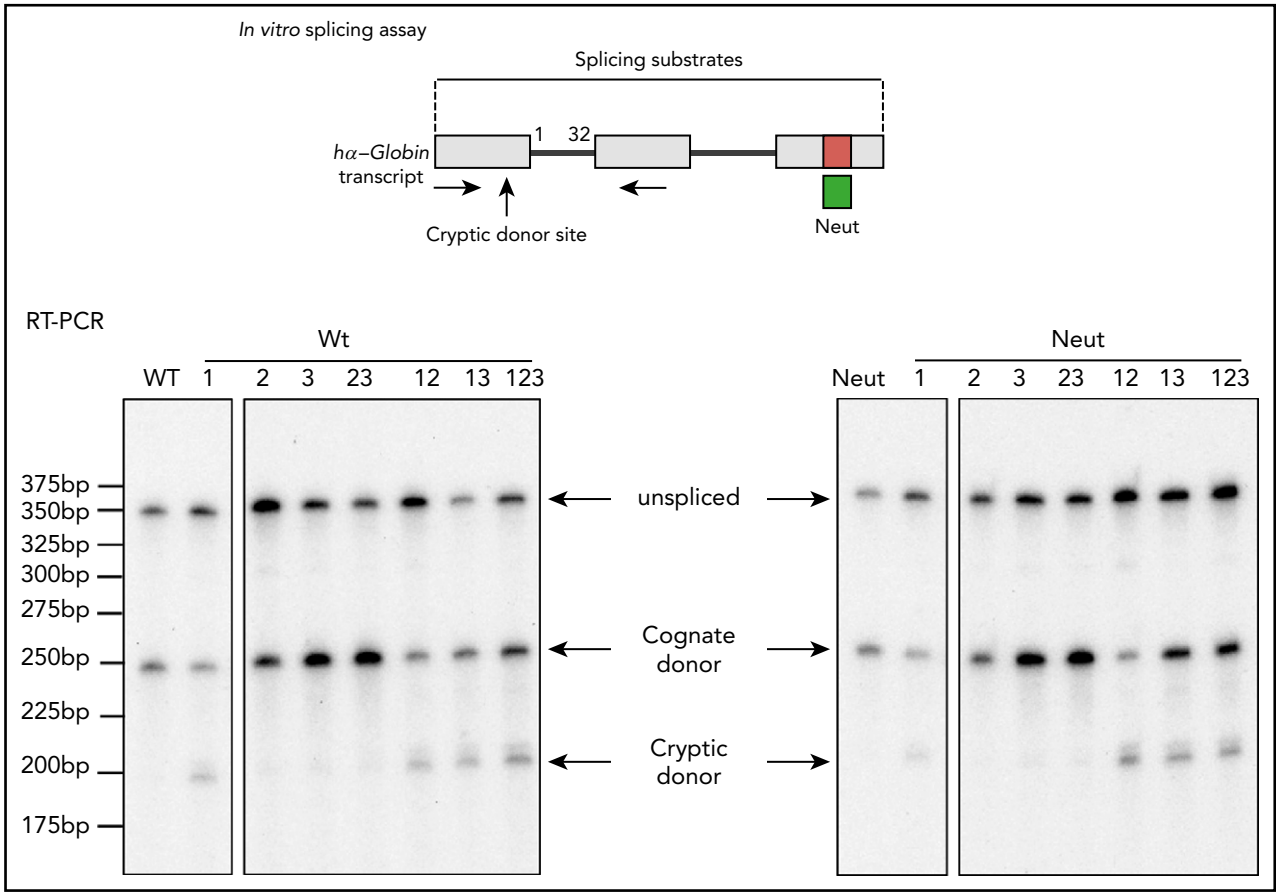

Figure 4. The splice regulatory function of the C-rich tract adjacent to the intron 1 splice donor is independent of the C-rich element within the 3′ UTR. The indicated set of intron 1 mutations (as in Figure 3) was assayed by in vitro splicing of a full-length WT *hα-globin* transcript with an intact 3′ UTR C-rich tract or with the C-rich tract substituted with a previously reported neutral (neut) sequence segment of the same size (green box).[7,8] Products of the splicing reaction were assessed by RT-PCR using the indicated amplimer set (horizontal arrows). The increase in cryptic RNA (cryptic donor) and reciprocal decrease in normally spliced *α-globin* RNA (cognate donor) in all substrates containing mutation 1 (Mut 1) confirm the regulatory activity of the C-rich tract 3′ to the intron 1 splice donor (Mut 1) on splice donor utilization. The lack of a similar impact by Muts 2, 3, or 23 or Neut further demonstrated that this activity of the donor site splice regulatory element is independent of the C-rich elements at the intron 1 splice acceptor and within the 3′ UTR. The 25-bp ladder size marker is indicated on the left.

When incubated in a mock-depleted extract, the splicing activity was restricted to the cognate splice donor (Figure 6 left panel). In contrast, depletion of PCBPs resulted in a marked shift of splicing to the cryptic donor site (Figure 6 right panel). These data lead us to conclude that the preferential use of the cognate intron 1 donor site is dependent on the interactions of the C-rich splice regulatory determinant with 1 or more polyC-binding proteins.

Discussion

Splicing of an RNA transcript can be constitutive or alternative. Constitutive splicing maintains the production of a single functional mRNA. The basis for constitutive splicing can be driven by strong cognate splice sites that effectively dominate the splicing pathway and/or may reflect the actions of *cis*-acting regulatory determinants that impart dominance of 1 splice site over another to maintain splicing fidelity. Here we demonstrate in the case of the *hα-globin* transcript that a C-rich determinant adjacent to the intron 1 donor site is critical to the expression of the *hα-globin* gene. Our study suggests that the C-rich splice regulatory determinant assembles an RNP complex that enforces functional splicing of the *hα-globin* transcript. This C-rich splice regulatory determinant therefore may constitute an essential determinant in the pathway of *hα-globin* gene expression.

The use of the cryptic splice donor site within exon 1 mRNAs may generate an out-of-frame RNA that would be a predicted target of the NMD pathway. Such instability would make it difficult to accurately quantify usage of this cryptic donor site relative to the normal cognate site in vivo. However, the analyses using in vitro splicing assays (Figures 3 and 4), in which nuclear to cytoplasmic transport, cytoplasmic translation, and the linked NMD pathway are not expected to play significant roles, suggest that the low levels of *hα-globin* mRNA generated from the cryptic donor site directly reflect a corresponding low activity of the cryptic donor site. This conclusion is further supported by the minigene analyses (Figure 5) in which the reporter RNA encoded by the 2-exon construct would not be subject to NMD. These data lead us to conclude that low levels of mRNA are generated from the cryptic donor in vivo (Figure 1A; supplemental Figure 1A) when it is situated in *cis* to the cognate donor.

In a prior transcriptome-wide analysis, we demonstrated that PCBPs affect alternative splicing of a defined subset of cassette exons in the human transcriptome that contain a C-rich

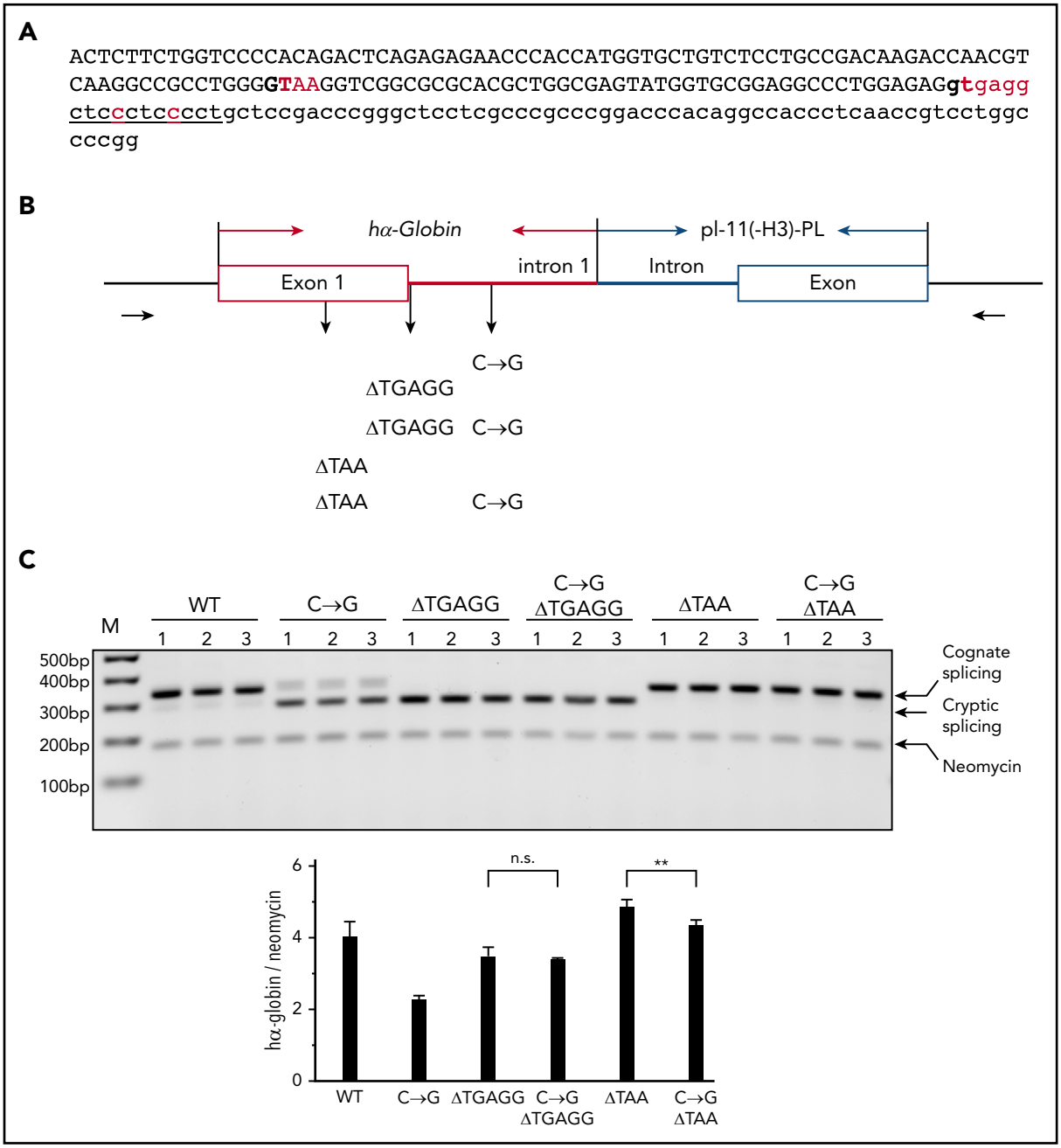

Figure 5. Confirmation of the importance of the C-rich motif 3′ to the intron 1 splice donor with regard to maintenance of cognate splicing activity (minigene assay). (A) The sequences of *hα-globin* gene exon 1 (upper case) and partial intron 1 (lower case) are shown. The GT dinucleotides that define the 2 splice donor sites are bolded. The cryptic splicing donor, the cognate splice donor, and the 2 Cs mutated to Gs are indicated in red font. (B) Schematic of the minigene constructs. C→G represents the 2 base substitutions in the C-rich sequence adjacent to the cognate splice donor (as in Figures 2 and 3). ΔTGAGG represents a 5-base naturally occurring α-thalassemia deletion that destroys the cognate splice donor.[29] ΔTAA is a 3-base deletion that destroys the cryptic splicing donor site within exon 1. The various combinations of these 3 mutations are displayed below the sequence. Each modified *hα-globin* sequence was linked to the downstream segment of the pl-11(-H3)-PL minigene vector as shown. The primer set used for RT-PCR analysis is indicated by black horizontal arrows below the schematic minigenes. (C) Each minigene was individually transfected into K562 cells in triplicate (indicated as 1, 2, and 3 above respective lanes). RT-PCR analysis of the cellular RNA was carried out 48 hours posttransfection to determine the impact of each mutation on the usage of the 2 competing splice donors. The chimeric minigene transcript mRNA and neomycin mRNA were coexpressed in *cis* from the minigene vector, and the *hα-globin*/*neomycin* mRNA ratio was calculated and is displayed in the histogram below the gel. M indicates 100-bp ladder size marker. Statistical significance (*P* values) was determined using 2-tailed, unpaired Student *t* test. $^{**}P < .01$. n.s., not significant.

polypyrimidine tract adjacent to their splice acceptor sites.[27] Of interest, those studies also revealed an enrichment of a C-rich motif adjacent to a subset of donor sites, the activities of which were enhanced by PCBPs.[27] The current study extends this second observation by directly demonstrating in the case of the *hα-globin* transcript that 1 or more polyC-binding proteins interact with a C-rich determinant adjacent to a splice donor site to enforce its predominant utilization over that of a competing cryptic donor site. Whether this mechanism is of more general importance can now be effectively explored.

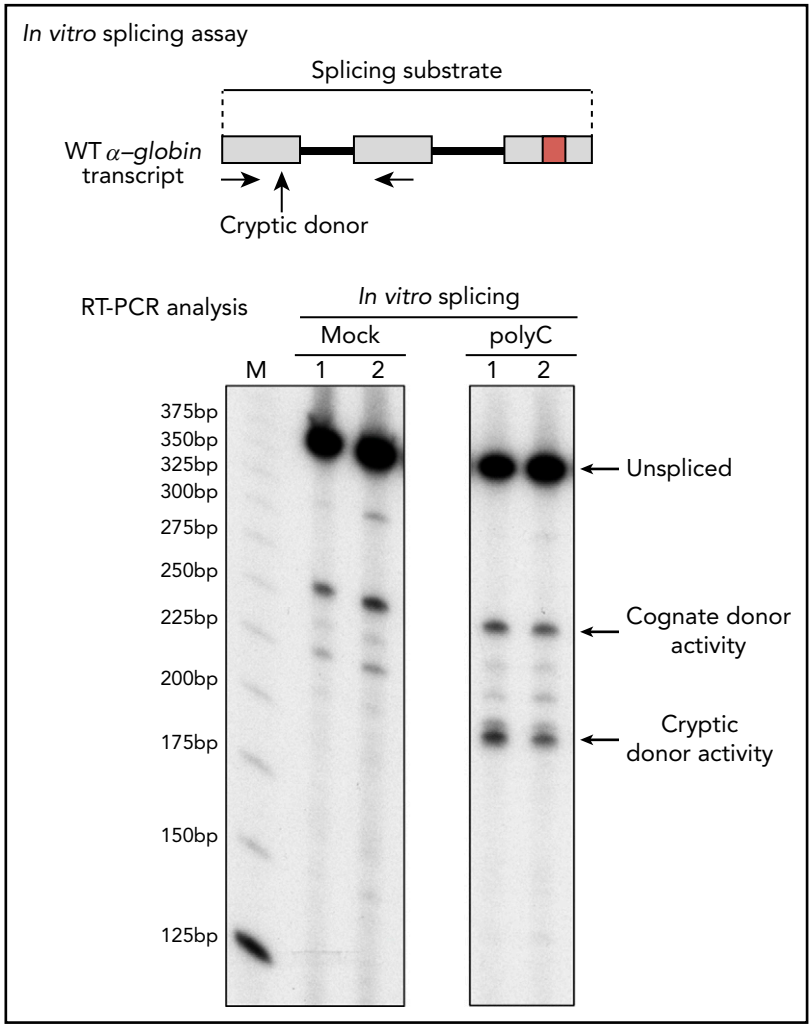

Figure 6. PCBPs enforce utilization of the cognate intron 1 splice donor. HeLa cell nuclear extracts were used for in vitro splicing studies after depletion of PCBPs (polyC) or mock depletion. The in vitro splicing of the full-length WT *α-globin* transcript was carried out in each of the 2 extracts (duplicate assays indicated as lanes 1 and 2). Comparison of the splicing products in the mock vs PCBP-depleted extract reveals the role of PCBPs in donor site selection and in enforcing functional splicing from the cognate splice donor. M indicates 25-bp ladder size marker.

We are left with the question of how the C-rich splice regulatory element in intron 1 enforces the predominant use of the cognate splice donor. It is of note that when the C-rich determinant is ablated, the cryptic site is favored over the cognate site (Figure 5). The impact of the C-rich element on the relative use of the 2 donor sites is unlikely to reflect their polarity or cotranscriptional regulatory mechanisms, because this element works as effectively in an in vitro splicing assay (Figures 3, 4, and 6) as it does in vivo (Figure 5). Therefore, one possible model is that the C-rich motif may act by preferentially enhancing assembly of the U1 small nuclear RNP complex at the cognate donor. Ongoing studies to directly to address this and other mechanistic models can now be pursued.

How do the current findings fit with the evolution of the *hα-globin* gene structure? The finding that a substantial fraction of the *hα-globin* transcript is shunted to a nonproductive splicing pathway in the absence of the C-rich determinant (Figures 3, 4, and 5) predicts that the fixation of the cryptic donor site in exon 1 in the human genome could only have occurred in the context of a coexisting (and neutralizing) C-rich splice regulatory determinant. This prediction is supported by the observation that the cryptic donor site (GT) within exon 1 is present in *cis* to a conserved C-rich splice control element in 8 of 11 primate species for which sequences are available (supplemental Figure 5A-B). In 3 remaining primate species (orangatan, gibbon, and bushbaby), the C-rich determinant is present in the absence of the cryptic donor (supplemental Figure 5A), but in no case is the cryptic donor present in the absence of the C-rich regulatory determinant. Therefore, although it remains unclear how the fixation of the cryptic splice donor in exon 1 might have imparted an evolutionary advantage so as to remain fixed in a majority of primate lineages, it seems that its appearance was most likely preceded by the C-rich splice regulator in intron 1. These evolutionary data further support the critical in vivo role of the C-rich splice regulator in *hα-globin* gene expression.

Acknowledgments

The authors thank Liebhaber laboratory members for sharing various reagents and thoughts.

This work was supported by National Institutes of Health, National Heart, Lung, and Blood Institute MERIT grant R01HL065449 (S.A.L.).

Authorship

Contribution: X.J. and S.A.L. conceptualized the study and designed the experiments; S.A.L. supervised the study; X.J. and J.H. performed the experimental work; and X.J. and S.A.L. wrote the paper.

Conflict-of-interest disclosure: The authors declare no competing financial interests.

Correspondence: Xinjun Ji, University of Pennsylvania, 415 Curie Blvd, Clinical Research Building Room 555, Philadelphia, PA 19104; e-mail: jixinjun@pennmedicine.upenn.edu.

Footnotes

Submitted 11 December 2018; accepted 26 February 2019. Prepublished online as *Blood* First Edition paper, 4 March 2019; DOI 10.1182/blood-2018-12-891408.

For original data, please contact Xinjun Ji at jixinjun@pennmedicine.upenn.edu.

The online version of this article contains a data supplement.

There is a *Blood* Commentary on this article in this issue.

The publication costs of this article were defrayed in part by page charge payment. Therefore, and solely to indicate this fact, this article is hereby marked "advertisement" in accordance with 18 USC section 1734.

REFERENCES

1. Moore MJ. From birth to death: the complex lives of eukaryotic mRNAs. *Science*. 2005; 309(5740):1514-1518.
2. Glisovic T, Bachorik JL, Yong J, Dreyfuss G. RNA-binding proteins and post-transcriptional gene regulation. *FEBS Lett*. 2008;582(14): 1977-1986.
3. Wang ET, Sandberg R, Luo S, et al. Alternative isoform regulation in human tissue transcriptomes. *Nature*. 2008;456(7221):470-476.
4. Pennsylvania State University. Globin Gene Server. http://globin.cse.psu.edu. Accessed 15 November 2018.
5. Thisted T, Lyakhov DL, Liebhaber SA. Optimized RNA targets of two closely related triple KH domain proteins, heterogeneous nuclear ribonucleoprotein K and alphaCP-2KL, suggest Distinct modes of RNA recognition. *J Biol Chem*. 2001;276(20): 17484-17496.
6. Holcik M, Liebhaber SA. Four highly stable eukaryotic mRNAs assemble 3′ untranslated region RNA-protein complexes sharing cis and trans components. *Proc Natl Acad Sci USA*. 1997;94(6):2410-2414.
7. Kong J, Ji X, Liebhaber SA. The KH-domain protein alpha CP has a direct role in mRNA stabilization independent of its cognate binding site. *Mol Cell Biol*. 2003;23(4): 1125-1134.
8. Kong J, Liebhaber SA. A cell type-restricted mRNA surveillance pathway triggered by ribosome extension into the 3′ untranslated region. *Nat Struct Mol Biol*. 2007;14(7): 670-676.
9. Ji X, Kong J, Carstens RP, Liebhaber SA. The 3′ untranslated region complex involved in stabilization of human alpha-globin mRNA assembles in the nucleus and serves an independent role as a splice enhancer. *Mol Cell Biol*. 2007;27(9):3290-3302.
10. Ji X, Kong J, Liebhaber SA. An RNA-protein complex links enhanced nuclear 3′ processing with cytoplasmic mRNA stabilization. *EMBO J*. 2011;30(13):2622-2633.
11. Weiss IM, Liebhaber SA. Erythroid cell-specific determinants of alpha-globin mRNA stability. *Mol Cell Biol*. 1994;14(12):8123-8132.
12. Weiss IM, Liebhaber SA. Erythroid cell-specific mRNA stability elements in the alpha 2-globin 3′ nontranslated region. *Mol Cell Biol*. 1995;15(5):2457-2465.
13. Kiledjian M, Wang X, Liebhaber SA. Identification of two KH domain proteins in the alpha-globin mRNP stability complex. *EMBO J*. 1995;14(17):4357-4364.
14. Wahl MC, Will CL, Lührmann R. The spliceosome: design principles of a dynamic RNP machine. *Cell*. 2009;136(4):701-718.
15. Haj Khelil A, Deguillien M, Morinière M, Ben Chibani J, Baklouti F. Cryptic splicing sites are differentially utilized in vivo. *FEBS J*. 2008; 275(6):1150-1162.
16. Kapustin Y, Chan E, Sarkar R, et al. Cryptic splice sites and split genes. *Nucleic Acids Res*. 2011;39(14):5837-5844.
17. Orkin SH, Goff SC, Hechtman RL. Mutation in an intervening sequence splice junction in man. *Proc Natl Acad Sci USA*. 1981;78(8): 5041-5045.
18. Bayat N, Farashi S, Hafezi-Nejad N, et al. Novel mutations responsible for α-thalassemia in Iranian families. *Hemoglobin*. 2013;37(2):148-159. 10.3109/03630269.2013.763821
19. Cürük MA, Baysal E, Gupta RB, Sharma S, Huisman TH. An IVS-I-117 (G→A) acceptor splice site mutation in the alpha 1-globin gene is a nondeletional alpha-thalassaemia-2 determinant in an Indian population. *Br J Haematol*. 1993;85(1):148-152.
20. Harteveld CL, Heister JG, Giordano PC, et al. An IVS1-116 (A→G) acceptor splice site mutation in the alpha 2 globin gene causing alpha + thalassaemia in two Dutch families. *Br J Haematol*. 1996;95(3):461-466.
21. Harteveld CL, Jebbink MC, van der Lely N, et al. Alpha-thalassemia phenotype induced by the new IVS-II-2 (T→A) splice donor site mutation on the alpha2-globin gene. *Hemoglobin*. 2006;30(1):3-7.
22. Pang W, Weng X, Ye X, et al. Identification of a variation in the IVSII of α2 gene and its frequency in the population of Guangxi. *Gene*. 2016;583(1):24-28.
23. Nishioka Y, Leder P. The complete sequence of a chromosomal mouse alpha–globin gene reveals elements conserved throughout vertebrate evolution. *Cell*. 1979;18(3):875-882.
24. Ji X, Kong J, Liebhaber SA. In vivo association of the stability control protein alphaCP with actively translating mRNAs. *Mol Cell Biol*. 2003;23(3):899-907.
25. Ji X, Wan J, Vishnu M, Xing Y, Liebhaber SA. αCP Poly(C) binding proteins act as global regulators of alternative polyadenylation. *Mol Cell Biol*. 2013;33(13):2560-2573.
26. Chkheidze AN, Lyakhov DL, Makeyev AV, Morales J, Kong J, Liebhaber SA. Assembly of the alpha-globin mRNA stability complex reflects binary interaction between the pyrimidine-rich 3′ untranslated region determinant and poly(C) binding protein alphaCP. *Mol Cell Biol*. 1999;19(7):4572-4581.
27. Ji X, Park JW, Bahrami-Samani E, et al. αCP binding to a cytosine-rich subset of polypyrimidine tracts drives a novel pathway of cassette exon splicing in the mammalian transcriptome. *Nucleic Acids Res*. 2016;44(5): 2283-2297.
28. Hovhannisyan RH, Carstens RP. A novel intronic cis element, ISE/ISS-3, regulates rat fibroblast growth factor receptor 2 splicing through activation of an upstream exon and repression of a downstream exon containing a noncanonical branch point sequence. *Mol Cell Biol*. 2005;25(1):250-263.
29. Felber BK, Orkin SH, Hamer DH. Abnormal RNA splicing causes one form of alpha thalassemia. *Cell*. 1982;29(3):895-902.