THE JOURNAL OF BIOLOGICAL CHEMISTRY
Vol. 276, No. 39, Issue of September 28, pp. 36337–36343, 2001
Printed in U.S.A.
© 2001 by The American Society for Biochemistry and Molecular Biology, Inc.

Characterization and Functional Implications of the RNA Binding Properties of Nuclear Factor TDP-43, a Novel Splicing Regulator of *CFTR* Exon 9*

Received for publication, May 10, 2001, and in revised form, July 17, 2001
Published, JBC Papers in Press, July 24, 2001, DOI 10.1074/jbc.M104236200

Emanuele Buratti and Francisco E. Baralle‡

From the International Center for Genetic Engineering and Biotechnology (ICGEB) 34012 Trieste, Italy

Variations in a polymorphic (TG)m sequence near exon 9 of the human *CFTR* gene have been associated with variable proportions of exon skipping and occurrence of disease. We have recently identified nuclear factor TDP-43 as a novel splicing regulator capable of binding to this element in the CFTR pre-mRNA and inhibiting recognition of the neighboring exon. In this study we report the dissection of the RNA binding properties of TDP-43 and their functional implications in relationship with the splicing process. Our results show that this protein contains two fully functional RNA recognition motif (RRM) domains with distinct RNA/DNA binding characteristics. Interestingly, TDP-43 can bind a minimum number of six UG (or TG) single-stranded dinucleotide stretches, and binding affinity increases with the number of repeats. In particular, the highly conserved Phe residues in the first RRM region play a key role in nucleic acid recognition.

We have recently reported the identification of TDP-43 as a splicing regulator that specifically binds the (UG)m-repeated polymorphic region near the 3′-splice site of *CFTR* exon 9 and down-regulates its recognition by the splicing machinery (1). This region, acting in concert with the adjacent (u)n element, is one of the key cis-acting sequences which regulate the proportion of exon 9 skipping in the mature CFTR mRNA transcript (1–3). Considering that exon 9 skipping produces a non-functional CFTR protein (4, 5) the study of the RNA binding properties of TDP-43 is of considerable importance to gain further insight concerning the potential disease-causing consequences of its binding *in vivo*. Indeed, the clinical relevance of these studies is highlighted by the existence of a clear association between certain (TG)m(T)n alleles with distinct forms of Cystic Fibrosis (1, 6–9).

In addition, the study of (UG)m elements can provide further insight concerning the mRNA splicing process in general because (UG)m sequences have been described to act as splicing regulatory sequences in different genomic contexts. In fact, in addition to the *CFTR* gene, the presence of simple (UG)m-repeated sequences has been described to influence the splicing process of at least two other genes: the apolipoprotein AII gene (10) and the human cardiac Na^+/Ca^{2+} exchanger (11). In the Apo AII gene the UG tract was shown to be functionally equivalent to a polypyrimidine tract and required for efficient splicing of Apo AII exon 2 (10) while in the human cardiac Na^+/Ca^{2+} exchanger (11) it acts as a strong intronic splicing enhancer situated in intron 2. It should be noted that in contrast with these two genes, the *CFTR* (UG)m element was found to possess a strong inhibitory effect on *CFTR* exon 9 splicing, a property that may probably be linked to its peculiar evolutionary history. In fact, sequencing of the mouse *CFTR* exon 9 genomic region has shown that in the flanking introns, the (TG)m(T)n regulatory elements are absent and that the intron themselves are of very different length when compared with the human introns (2). This finding, together with the observation that mouse *CFTR* exon 9 is not subject to alternative splicing, suggests that the presence in humans of the (UG)m sequence represents a disturbing element, which interferes with the normal maturation process of the CFTR pre-mRNA. This conclusion is also supported by the fact that *CFTR* exon 9 and its intronic flanking sequences are found co-integrated with characteristic L1 sequences in multiple chromosome locations distinct from the *CFTR* locus (12, 13). These findings may indicate that the introduction of foreign elements in the CFTR IVS8 and IVS9 sequences may be a consequence of a retrotransposition event, which affected the human *CFTR* gene early during the course of evolution.

In order to better elucidate the role of the CFTR (UG)m element and obtain functional clues regarding the role of TDP-43 in the splicing process we have characterized the RNA/DNA binding properties of TDP-43. Our results have confirmed the existence in this protein of two fully functional RNA recognition motifs (RRM),[1] also known as RBD, for RNA binding domains (14–18), which possess distinct binding characteristics.

EXPERIMENTAL PROCEDURES

Plasmid Construction and Oligonucleotides—Plasmids pTCTT3 and pTG12 and minigenes lacking the (TG)m/T(n) elements were obtained as previously described (1). Plasmids pTG3, pTG6, pTG9 were obtained by annealing the following forward and reverse oligos and ligating them in pBluescript KS (Stratagene) linearized with *Sma*I: 5′-gaaaattaatgtgtggaaaattaagaaa-3′ (oligo TG3) and 5′-tttcttaattttccacacattaattttc-3′ (oligo AC3) for pTG3, 5′-gaaaattaatgtgtgtgtgtggaaaatt-3′ (oligo TG6) and 5′-aattttccacacacacattaattttc-3′ (oligo AC6) for pTG6, 5′-gaaaattaatgtgtgtgtgtgtgtgtga-3′ (oligo TG9) and 5′-tcacacacacacacacattaattttc-3′ (oligo AC9) for pTG9. The plasmid pTAR was obtained by annealing the following primers and ligating them in pBluescript KS (Stratagene) linearized with *Sma*I: 5′-ctgctttttgcctgtactgggtctctctggttagaccagatctgag-3′ (oligo TARS) as the forward and 5′-ctcagatctgtctaaccagagagacccagtacaggcaaaaagcag-3′ (oligo TARAS) as the reverse primer. The synthetic $(UG)_{12}$ oligo was obtained from MWG Biotech (Firenze, Italy).

Expression of Recombinant TDP-43 as a GST Fusion Protein—The

* This work was supported by Telethon Onlus Foundation Grant E1038. The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

‡ To whom correspondence should be addressed: Padriciano 99, 34012 Trieste, Italy. Tel.: 0039-40-3757337; Fax: 0039-40-3757361; E-mail: baralle@icgeb.trieste.it.

[1] The abbreviations used are: RRM, RNA recognition motif; GST, glutathione *S*-transferase; PAGE, polyacrylamide gel electrophoresis; EMSA, electromobility shift assay; SSR, short sequence repeats.

This is an Open Access article under the CC BY license.

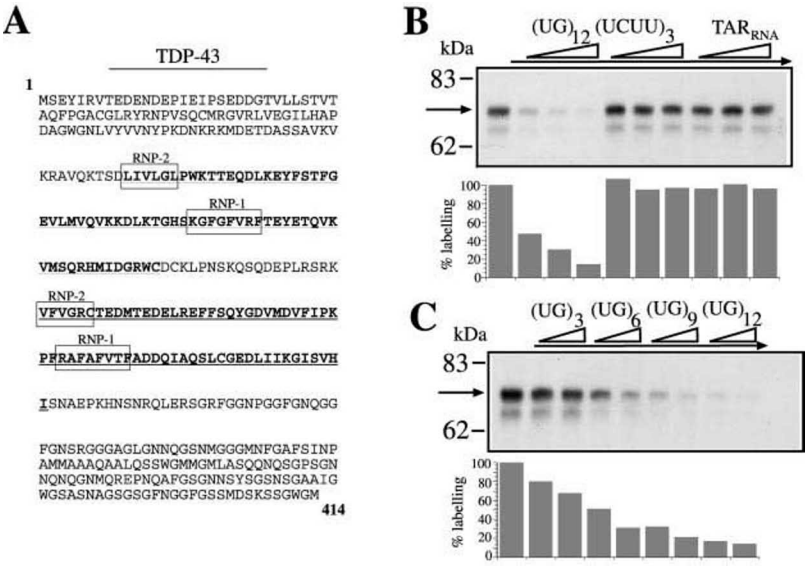

FIG. 1. **RNA binding specificities of TDP-43.** *A,* the amino acid sequence of TDP-43 (residues 1–414) with the two predicted RRM motifs (residues 106–175 and 193–257) highlighted in *bold* (RRMs were identified with a search using the Pfam program at www.sanger.ac.uk). The boxed regions highlight the highly conserved RNP-2 and RNP-1 regions. *B,* UV cross-linking competition analysis loaded on a 10% SDS-PAGE gel of GST-TDP43 bound to labeled $(UG)_{12}$ RNA in the presence of increasing amounts of cold $(UG)_{12}$, $(UCUU)_3$, and TAR RNA (the excess molar ratios of cold competitor RNA used in each data point was 3, 8, and 15). The *lower panel* shows a graph with the percentage of TDP-43 labeling following incubation with the cold competitors. *C,* competition analysis using short RNAs of equal length containing different numbers of (UG) repeats: $(UG)_3$, $(UG)_6$, $(UG)_9$, and $(UG)_{12}$. The molar ratio of cold competitor RNA to labeled RNA used for each data point was 8 and 15. The *lower panel* shows a graph with the percentage of TDP-43 labeling following incubation with the cold competitors.

GST-TDP43 and GST-TDP43(101–261) fusion proteins were obtained as previously described (1). Deletion of the RRM1, RRM2, and 106–111 RNP-2 regions was obtained using a two step polymerase chain reaction extension method with sense and reverse primers spanning RRM1 (5′-aaaacatccgataaacttcctaat-3′ and 5′-attaggaagtttatcggatgtttt-3′), RRM2 (5′-ttgagaagcagatccaatgccgaa-3′ and 5′-ttcggcattggatctgcttctcaa-3′), and 106–111 RNP-2 (5′-aaaacatccgatccatggaaaaca-3′ and 5′-tgtttttccatggatcggatgtttt-3′). It should be noted that these regions were identified through a search using the Pfam program at www.sanger.ac.uk. In order to introduce the L106D, V108D, and L111D mutations we used the following primers: 5′-tccgatgatatagatttgggtgatccatgg-3′ and 5′-ccatggatcacccaaatctatatcatcgga-3′. Similarly, the Phe residues (at positions 147 and 149) in the 145–152 RNP-1 motif were mutated to Leu using the following forward and reverse primers: 5′-aaggggttgggcttggttcgtttt-3′ and 5′-aaaacgaaccaagcccaacccctt-3′. The single Phe-194 in the 193–197 RNP-2 motif was mutated to Leu using the following forward and reverse primers: 5′-gtgttggtggggcgctgt-3′ and 5′-acagcgccccaccaacac-3′. The two Phe residues (at positions 229 and 231) in the 227–234 RNP-1 motif were mutated to Leu using the following forward and reverse primers: 5′-agggccttggccttggttacattt-3′ and 5′-aaatgtaaccaaggccaaggccct-3′. Double and triple mutants were obtained using the same methodology on single- and double-mutated proteins. All fusion proteins were expressed in *Escherichia coli* DH5α cells by overnight induction at room temperature in the presence of 0.1–0.3 mM IPTG. Cells were then resuspended in phosphate-buffered saline, 1% Triton X-100, and sonicated. The supernatant was recovered after centrifugation at $3000 \times g$ for 30 s in an Eppendorf 5810R centrifuge and incubated with glutathione *S*-Sepharose 4B beads (Amersham Pharmacia Biotech). The absorbed proteins were then eluted according to the manufacturer's instructions. Purified proteins were quantitated on an SDS-PAGE gel using bovine serum albumin standards (Sigma).

UV Cross-linking Assay—Plasmids were linearized by digestion with *Hin*dIII and transcription was performed with T7 RNA polymerase (Stratagene) in the presence of labeled [α-^{32}P]UTP, DNase-treated according to standard protocols, and purified on a Nick column (Amersham Pharmacia Biotech) according to the manufacturer's instructions. The labeled RNAs were then precipitated and resuspended in RNase-free water. The UV cross-linking assay was performed by adding [α-^{32}P]UTP-labeled RNA probes (1×10^6 cpm per incubation) in a water bath for 15 min at 30 ° C with 200 ng of each different purified protein in a 20-μl final volume. Binding conditions were 20 mM Hepes pH 7.9, 72 mM KCl, 1.5 mM $MgCl_2$, 0.78 mM magnesium acetate, 0.52 mM dithiothreitol, 3.8% glycerol, 0.75 mM ATP, and 1 mM GTP. In the competition experiments, cold RNA and DNA were also added as competitors 5 min before addition of the labeled RNAs (the molar excess of the unlabeled competitor used in the different experiments is stated in each figure legend). Samples were then transferred in the wells of an HLA plate (Nunc, InterMed) and irradiated with UV light on ice (0.8 joules, ~5 min) using a UV Linker (Euroclone). Unbound RNA was then digested with 30 μg of RNase A (Sigma) and 6 units of RNase T1 (Sigma) by incubating at 37 °C for 30 min in a water bath and then adding SDS-PAGE sample buffer. Samples were then analyzed on a 10% SDS-PAGE gel followed by autoradiography with autoradiographic XAR film (Kodak). Films were then scanned on a Macintosh G3 work station using Adobe Photoshop and printed using a Phaser 400 printer.

Electromobility Shift Assay (EMSA)—Oligonucleotides (200 ng, ~25 pmols) were labeled by phosphorylation with [γ-^{32}P]ATP and T4 polynucleotide kinase (PNK, Stratagene) for 1 h at 37 °C and then precipitated in 0.3 M sodium acetate, pH 5.2 and three volumes of ethanol. After centrifugation and a washing step with 70% ethanol the labeled oligos were resuspended in 400 μl of water. Each binding reaction was performed at room temperature for 15 min by mixing the purified protein with the labeled oligo (or RNA) in a 20-μl final volume. The reactions were performed in 1× bind shift binding buffer (20 mM Hepes pH 7.9, 2 mM $MgCl_2$) and electrophoresed on a 5% polyacrylamide gel at 100 V for 1 h in 0.5× Tris borate/EDTA buffer at 4 °C. The gel was then dried on 3 MM Whatman filter paper and exposed for 20 min with autoradiographic XAR film (Kodak). For quantitation gels were measured with an InstantImager (Packard).

RESULTS

Binding Specificity of TDP-43 for Different RNA Sequences—In our search to identify proteins that recognize the splicing regulatory elements of *CFTR* exon 9, we recently isolated TDP-43 as a protein that binds specifically to the splicing regulatory (UG)m element found near the 3′-splice site of this exon (1). Up to now, the only other described cellular function of TDP-43 was the ability to bind a HIV-1 TAR DNA polypyrimidinic sequence motif leading to the inhibition of HIV-1 transcription (19). Interestingly, no binding to TAR RNA had been reported (19). However, our recent observation that TDP-43 can efficiently bind to (UG)m sequences (1) is consistent with the presence of two putative full-length RRM domains (Fig. 1*A*) located between residues 106 and 175 (RRM1) and 193 and 257 (RRM2) of its coding sequence according to the output of the Pfam program (available at www.sanger.ac.uk). This finding provided a functional basis that accounted for the ability of

TABLE I
ssDNA oligonucleotides used for competition analysis

Nucleotide sequence (5′-3′)	Oligo
gaaaattaatgtgtggaaaattaagaaa	TG3
gaaaattaatgtgtgtgtgtggaaaatt	TG6
gaaaattaatgtgtgtgtgtgtgtgtgtga	TG9
tgtgtgtgtgtgtgtgtgtgtgtg	TG12
tttcttaattttccacacattaattttc	AC3
aattttccacacacacattaattttc	AC6
tcacacacacacacacacattaattttc	AC9
acacacacacacacacacacacac	AC12
tcctcctccttcttcttcttcagg	TCTTS
cctgaagaagaagaaggaggagga	TCTTAS
ctgctttttgcctgtactgggtctctctggttagaccagatctgag	TARS
ctcagatctggtctaaccagagagacccagtacaggcaaaaagcag	TARAS
gAAAATTAACAATTTAAA	mEX9AS

TDP-43 to bind RNA sequences, and in this study we provide a detailed analysis of their functionality and importance.

Initially, using a GST fusion protein containing the TDP-43 full coding sequence (GST-TDP43), we confirmed the binding specificities of recombinant GST-TDP43 toward different RNAs in UV-cross-linking analysis. Fig. 1*B* shows that increasing amounts of unlabeled $(UG)_{12}$ RNA were very efficient competitors. Because (UG)m sequences have been described as efficient polypyrimidinic sequences (10) we also tested the possibility that recognition could be extended generally to this type of sequences. However, addition of a cold polypyrimidinic RNA $(UCUU)_3$ did not have any effect on the binding of GST-TDP43 to $(UG)_{12}$ (Fig. 1*B*), confirming the high specificity of TDP-43 binding to UG-repeated motifs. This result is consistent with our previously reported pull-down assay, which did not result in any TDP-43 being recognized by $(UCUU)_3$ RNA (1). The high sequence binding specificity of TDP-43 is also highlighted by the fact that addition of cold TAR RNA was incapable of competing with the binding of GST-TDP43 to the $(UG)_{12}$ sequence, a finding that had already been described in the original isolation of TDP-43 (19) and which we confirm here. It should be noted that at present no other RNA-binding protein has been described to bind UG-repeated sequences although CUG-BP (CUG-binding protein) has been recently described to bind UG repeats in a yeast three-hybrid system (20).

It was then of interest to analyze the minimum length of (UG) repeats that could be specifically bound by this protein. Therefore, cold RNAs competitors containing different lengths of (UG) repeats (3, 6, 9, and 12) were incubated in the presence of GST-TDP43 and labeled $(UG)_{12}$ RNA. The results shown in Fig. 1*C* demonstrate that efficient competition can be observed only when the number of (UG) repeats is equal or above six and that there is a relationship between the number of (UG) repeats and the efficiency of binding.

The DNA Binding Specificity of TDP-43 Includes Single-stranded (TG)-repeated Sequences but Not Double-stranded (tg/ac)-repeated Sequences—The fact that this protein was originally described to bind TAR DNA sequences raised the possibility that the binding characteristics of this protein might include (TG)-repeated sequences as well as (UG)-repeated sequences. Therefore, we performed competition analysis using GST-TDP43 bound to $(UG)_{12}$ RNA in the presence of cold single-stranded DNA oligos (see Table I). As shown in Fig. 2*A* the most efficient competitor was represented by the oligo TG12 carrying twelve (TG) repeats followed by the TARS oligo, a result that is consistent with what had been observed by Ou *et al.* (19). It should also be noted that other oligos (in particular TCTTS) display a weak but significant ability to compete for this protein, a characteristic not found in the $(UCUU)_3$ RNA (see Fig. 1). This finding also confirms the original observations of Ou *et al.* (19) who, using a polymerase chain reaction-based

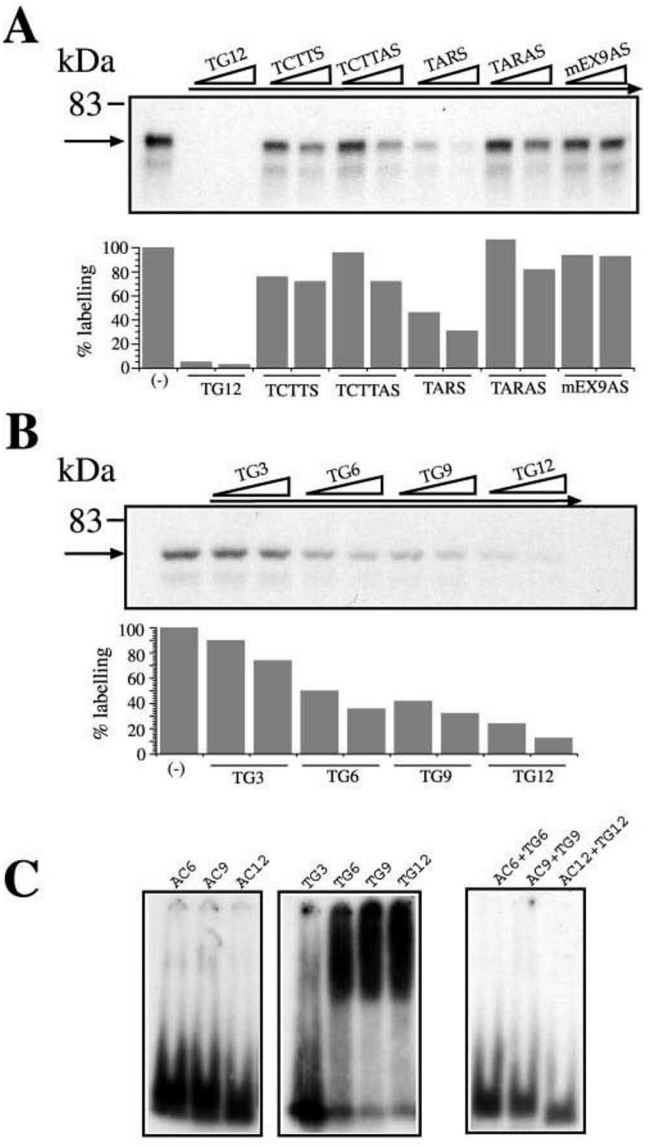

FIG. 2. **DNA binding specificities of TDP-43.** *A,* competitive ability of different single-stranded DNA oligos on the binding of GST-TDP43 to labeled $(UG)_{12}$. For each oligo, the molar ratio of cold competitor DNA to labeled $(UG)_{12}$ RNA used for the two data points was 5 and 10. The *lower panel* shows a graph with the percentage of TDP-43 labeling following incubation with the cold oligos. *B,* competition analysis of the effects of single-stranded oligos containing an increasing number of (TG) repeats (3, 6, 9, and 12) on the binding of GST-TDP43 to labeled $(UG)_{12}$. For each oligo the molar ratio of competitor cold DNA to labeled $(UG)_{12}$ RNA used for the two data points was 5 and 10. The *lower panel* shows a graph with the percentage of GST-TDP43 labeling following incubation with the cold oligos. *C,* EMSA analysis of GST-TDP43 binding to different 5′-labeled single-stranded (AC)-rich oligos (AC6-AC12, *left panel*) as opposed to analogous (TG)-rich oligos (TG3-TG12, *middle panel*). *C, right panel,* EMSA analysis using the double-stranded oligos (labeled only on the (AC)-containing strand) with GST-TDP43. Complexes were fractionated on a 5% non-denaturing polyacrylamide gel.

site selection procedure, found that recombinant TDP-43 preferably bound DNA stretches of eight contiguous pyrimidine residues. Nonetheless, Fig. 2*B* shows that the use of oligos containing different numbers of TG repeats yielded results very similar to those obtained in Fig. 1*C* using (UG)-repeated sequences. Also in this case, the minimum number of (TG) repeats needed to efficiently compete for GST-TDP43 binding is six, and there is a relationship between the number of (TG) repeats and the efficiency of competition.

Finally, oligos containing (AC) repeats can function as effi-

cient competitors for the binding of GST-TDP43 to $(UG)_{12}$ (data not shown). However, in this case competition was caused by the (AC) repeats annealing directly to the $(UG)_{12}$-labeled sequence and inhibiting binding of the protein rather than by binding directly to GST-TDP43. In fact, EMSA analysis shows that there is little if any direct binding of GST-TDP43 to AC6, AC9, or AC12 end-labeled oligos (Fig. 2*C*, *left panel*) while direct binding of GST-TDP43 efficiently occurs for single-stranded (TG)-repeated sequences containing 6, 9, and 12 (TG) repeats (Fig. 2*C*, *central panel*). Notably, the fact that the oligo bearing three tg-repeats (TG3) could not efficiently bind GST-TDP43 confirms the previous competition data by UV-cross-linking (Fig. 2*B*). Thus, in order to establish whether TDP-43 could bind double-stranded oligos we then annealed labeled (AC) oligos with equal amounts of complementary and unlabeled (TG) oligos and then repeated the EMSA analysis. The results confirm that double-stranded oligos containing TG repeats do not bind TDP-43 (Fig. 2*C*, *right panel*).

Comparing the (TG) and (UG) Binding Efficiencies of TDP-43; Formation of Two Distinct Complexes with ssDNA as Opposed to Only One with ssRNA in a UV Cross-linking Assay—The DNA and RNA binding efficiencies of TDP-43 were assumed to depend on the two RRM regions. In order to establish whether there was no other protein domain involved in RNA/DNA recognition and to compare the two binding efficiencies we produced a construct coding for a truncated TDP-43 protein lacking the N- and C-terminal regions (Fig. 3*A*). The DNA and RNA binding efficiencies of GST-TDP43 and GST-TDP43-(101–261) were then compared using as substrate a 5′-labeled $(TG)_{12}$ or $(UG)_{12}$ oligo (at a fixed concentration of 6 nM). The results, shown in Fig. 3*B*, demonstrate that deletion of the N- and C-terminal regions of TDP-43 does not appear to affect the RNA binding efficiency of the central RRM-containing region and may even slightly enhance it. It should be noted that the retarded complexes formed by each protein do not migrate at the same level, an indication of the higher molecular weight of the GST-TDP43/nucleic acid complex as opposed to the GST-TDP43(101–261)/nucleic acid complex.

Interestingly, binding of TDP-43 to $(UG)_{12}$ as opposed to $(TG)_{12}$ presents some differences as well. In fact, Fig. 3*C* shows that in UV cross-linking analysis only one RNA-protein complex of 50–52 kDa is formed when GST-TDP43-(101–261) is bound to $(UG)_{12}$, as previously described (1). On the other hand, at least two major DNA-protein complexes with altered mobility can be detected when the same protein is bound to $(TG)_{12}$. The formation of multiple complexes in SDS-PAGE following UV cross-linking analysis has already been described for another RRM protein, Gbp1p (21), as a result of the formation of multiple covalent cross-links between protein and nucleic acid. The fact that the migrating complexes are different when using $(UG)_{12}$ as opposed to $(TG)_{12}$ represents a further indication that TDP-43 RNA and DNA binding characteristics may not be identical.

Binding of TDP-43 Derivatives Lacking the RRM1 or RRM2 Motifs—In order to test the importance of each TDP-43 RRMs we then made a series of deletion mutants and analyzed their binding to $(UG)_{12}$ RNA. Fig. 4*A* shows a schematic representation of two mutants in which we selectively deleted the first RRM motif (GST-TDP43/ΔRRM1) and the second RRM motif (GST-TDP43/ΔRRM2). In addition, in order to confirm the presence of the first RNP-2 motif (residues 106–111), which had not been previously detected in the original work by Ou *et al.* (19) we prepared two mutants, the first containing a deletion of the entire 6-amino acid region (GST-TDP43/Δ(106–111)) and the second introducing an Asp residue in substitution for the three amino acids (Leu-106, Val-108, and Leu-111), which

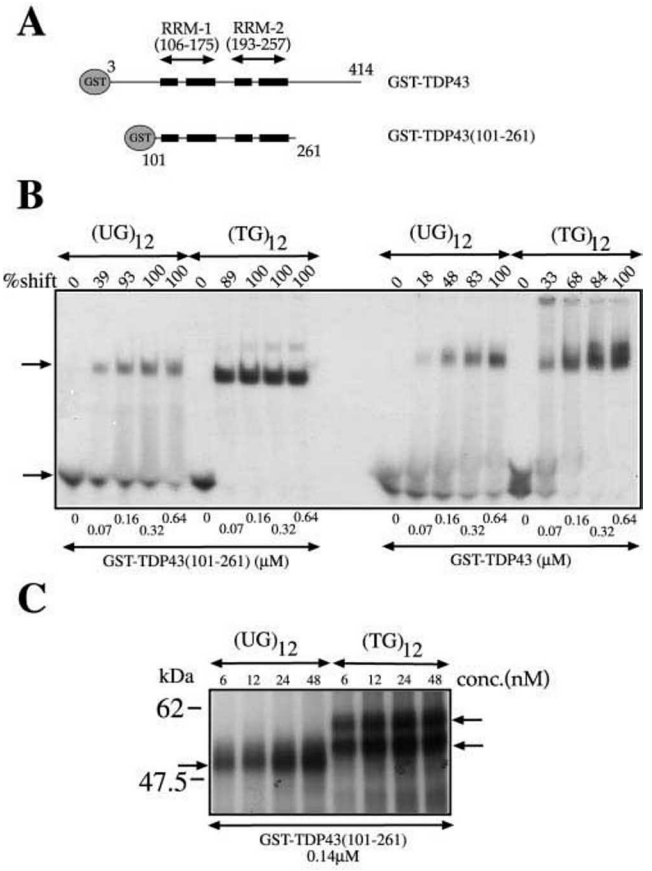

FIG. 3. **Comparison of ug/tg binding properties of TDP-43.** *A,* schematic representation of the GST-TDP43 protein and of the truncated mutant GST-TDP43-(101–261). *B,* reactivity in EMSA analysis of both proteins with labeled single-stranded $(UG)_{12}$ RNA (6 nM) or labeled $(TG)_{12}$ DNA (6 nM). Protein concentrations ranged from 0.07 μM to 0.64 μM. The *arrows* on the *left* indicate the retarded protein-nucleic acid complexes (*upper arrow*) and the free nucleic acid (*lower arrow*). The percent shift for each data point as quantified using a phosphorimager is indicated. *B,* reactivity of UV cross-linking of GST-TDP43-(101–261) protein (0.14 μM) with labeled single-stranded $(UG)_{12}$ RNA or labeled $(TG)_{12}$ DNA at different concentrations (6–48 nM). The *arrows* on the *left* indicate the retarded protein-RNA complex while the *arrows* on the *right* indicate the retarded protein-DNA complexes.

were predominantly conserved in the corresponding RNP-2 motifs of well characterized RNA-binding proteins (see Fig. 5). The four mutants were then analyzed by EMSA analysis using labeled $(UG)_{12}$ RNA. The results show that deletion of RRM1 and deletion (or mutation) of the 106–111 RNP-2 motif completely abolished the ability of TDP-43 to bind the RNA (Fig. 4*B*, *first and second upper panels*). This result not only confirms the presence of a functional RNP-2 motif localized in position 106–111 but also suggests that the RRM1 sequence spanning residues 106–175 is of fundamental importance for the binding to RNA.

It should be noted that deletion of RRM2 (leaving the RRM1 sequence intact) does not completely abolish the RNA binding capability of TDP-43 but leads to the appearance of a super-shifted RNA-protein complex (Fig. 4*B*, *first panel*), suggesting that the binding characteristics of RRM2 are quite different from those of RRM1. The specificity of this complex formation is confirmed by a competition analysis (Fig. 4*C*) in which we added increasing amounts of each mutant (GST-TDP43/ΔRRM1 and GST-TDP43/ΔRRM2) to a reaction mix that contained GST-TDP43 and labeled $(UG)_{12}$ RNA. The results show that the GST-TDP43/ΔRRM2 mutant is capable of actively competing with GST-TDP43 for the formation of the super-

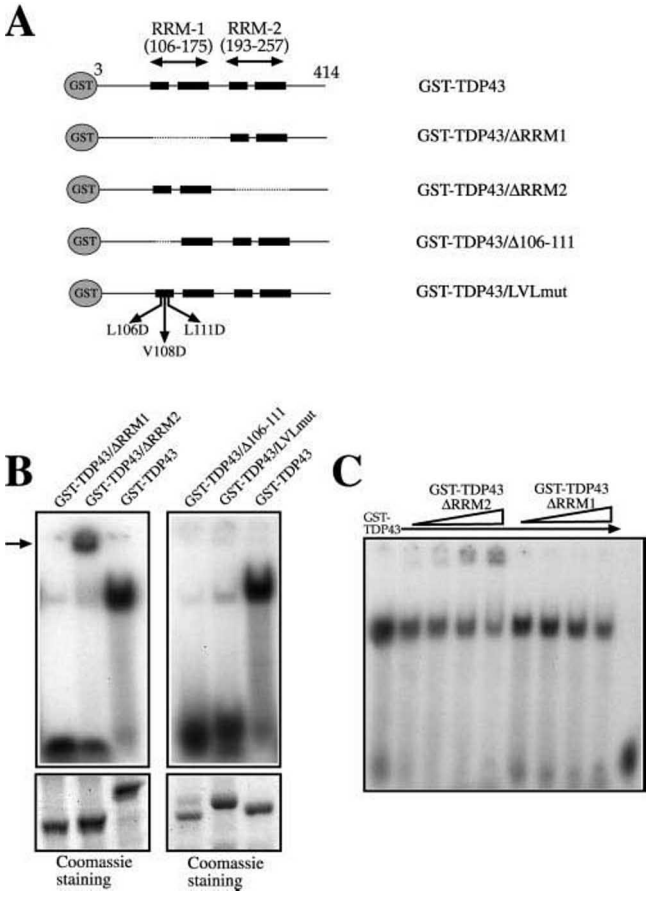

FIG. 4. **Deletion and mutational analysis of TDP-43 RRM regions and their effects on RNA binding ability.** *A,* schematic panel of the deletions and mutations introduced in the different RRM motifs of the GST-TDP43 protein. *B*, *first* and *second upper panels*, EMSA analysis of the effects of these mutations on the binding to 5′-labeled single-stranded $(UG)_{12}$ RNA and each mutant protein. The *arrow* on the *left* indicates the anomalous retarded protein-nucleic acid complex that is observed in the case of GST-TDP43/ΔRRM2 incubated with $(UG)_{12}$ RNA. *B*, *lower panels*, two SDS-PAGE gels stained with Coomassie Blue for each protein used in the EMSA analysis. *C,* competition analysis performed by adding increasing molar quantities (1, 2, 4, and 8, respectively) of GST-TDP43/ΔRRM2 and GST-TDP43/ΔRRM1 to a saturated GST-TDP43/$(UG)_{12}$ complex. The *last lane* contains $(UG)_{12}$ alone as control.

shifted complex, but no change is observed following addition with GST-TDP43/ΔRRM1.

Importance of Conserved Aromatic Residues in TDP-43 RRM Domains for RNA Binding—In order to better characterize how TDP-43 binds to RNA we compared the sequence of TDP-43 RRMs domains with that of other well characterized RRMs found in proteins whose structure has been solved by crystallography: hnRNP A1 (22), Sxl (23), PABP (24) and U1A spliceosomal protein (18, 25). Fig. 5 shows the RRM domains found in these proteins, which are most similar to RRM1 and RRM2 of TDP-43. It should be noted that very little homology was detected between U1A RRMs and TDP-43 RRMs (data not shown). Overall, the highest amino acid identity between the different RRMs can be found in correspondence with the highly conserved RNP-1 and RNP-2 consensus motifs. In particular, several key aromatic residues that have been reported to be responsible for direct stacking interactions with RNA bases in these different proteins (marked with *open circles*) are conserved in TDP-43 RRMs. The only exception is represented by the first putative TDP-43 RNP-2 motif (residues 106–111) in which none of the aromatic residues reported to make direct stacking interactions with the RNA are conserved. It should be noted that mutation of these aromatic residues in hnRNPA1 (26) and U1A (27, 28) has long been known to severely affect the RNA binding capability of these proteins. To further investigate these similarities we then prepared a series of GST-TDP43-(101–261) mutants in which the conserved Phe residues in each RNP motif were mutated to leucine residues (Fig. 6*A*). The rationale for this change in residue resides in the fact that a Phe to Leu single amino acid mutation has been previously described to abolish the functionality of the RNP motif in the case of the nucleolin protein (29). Each mutant was expressed in *E. coli* (Fig. 6*B*), and its ability to bind $(UG)_{12}$ RNA and $(TG)_{12}$ DNA was assayed by EMSA (Fig. 6, *C* and *D*). In both cases, the mutations that most reduced binding to the nucleic acid were the F147L and F149L in the RNP-1 motif of RRM1. The importance of these residues is best reflected in the fact that in double mutants 5 and 7 the ability to bind $(UG)_{12}$ RNA and $(TG)_{12}$ DNA was lost while in the double mutation that preserved intact only Phe-147 and Phe-149 (mutant number 6) binding could still occur. This result is in accordance with the results obtained by Ou *et al.* (19), who performed progressive deletions of TDP-43 and observed that ability to bind TAR DNA was lost only when the first RNP-1 motif was deleted. This observation was also confirmed by incubating labeled TAR DNA oligo with our mutants, an experiment that yielded identical results to those obtained in Fig. 6*D* for the TG12 oligonucleotide (data not shown).

Finally, it should be noted that while deletion of the entire RRM2 domain leads to a supershifted complex (Fig. 4*B*) the point mutations of the aromatic residues in RRM2 (Fig. 6*C, mutant 6*) result in a complex whose mobility is indistinguishable from the wild type. Taken together, these results suggest the presence of considerable interplay between the RRM1 and RRM2 domains of TDP-43.

DISCUSSION

The human genome has been recently shown to be heavily composed of repetitive elements (>50%) that vary in complexity from whole genes and very long stretches of DNA to much simpler and shorter nucleotide sequences (30, 31). These shorter sequences are commonly known as short sequence repeats (SSRs) and are often highly polymorphic (30). In particular, SSR elements are estimated to contribute 3% of the whole genome (with simple dinucleotide repeats accounting alone for 0.5% of the total) (31). Usually, SSRs are composed of repeated nucleotide motifs ranging from 1 to 20 nucleotides in length and are present in blocks of up to thousands of tandem units (30, 32). Although their function is still largely unknown repetitive nucleotide stretches are known to play important roles in several pathological conditions. For example, expansion of simple trinucleotide repeats through a mechanism of dynamic mutation is known to cause distinct human genetic diseases such as myotonic dystrophy (33), the Fragile X Syndrome (34, 35), Huntington's disease (36), or a series of neurodegenerative diseases (37). Moreover, SSR sequences have also been described to affect the splicing process of the *CFTR* gene and correlate with severity of disease (1, 6–9).

In this study we report the characterization of the novel RNA/DNA binding properties of TDP-43 (19), a protein that we have recently described to play a role in *CFTR* exon 9 splicing and occurrence of disease following binding to the (UG)m regulatory sequence (1). Interestingly, we have found that TDP-43 can efficiently bind a number of (UG) repeats equal to or greater than six and that there is a relationship between the number of (UG) repeats and the efficiency of binding. This finding provides a functional explanation for our recent demonstration of a connection between the length of the (UG)m regulatory region and alternative splicing of *CFTR* exon 9 (1).

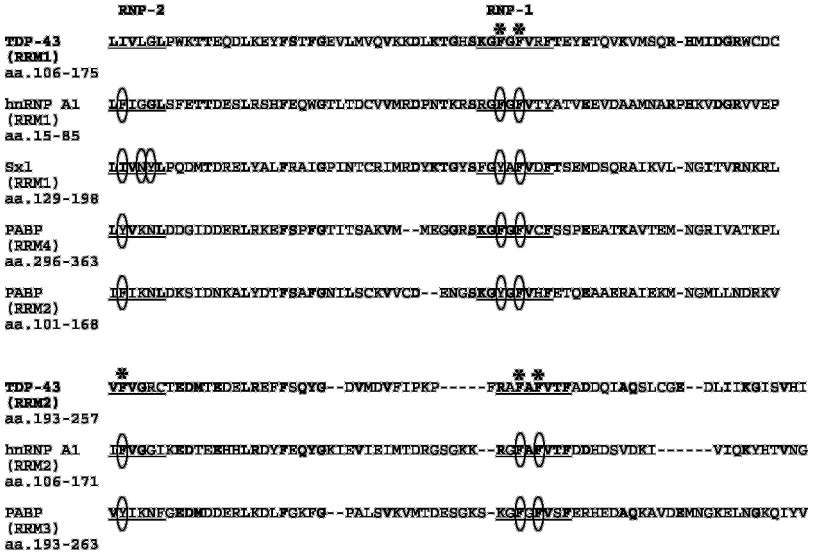

FIG. 5. **Comparison of TDP-43 RRM regions with similar RRM motifs of known tertiary structure.** This figure shows a comparison of the two TDP-43 RRM motifs with homologous RRM motifs belonging to proteins for which their crystallographic structure is known, hnRNP A1, Sxl, and PABP. Sequence identities with the TDP-43 sequence are marked in *bold lettering* while key residues involved in direct stacking interactions with the RNA, as confirmed by structural analysis are highlighted with *open circles*. The highly conserved RNP-1 and RNP-2 consensus sequences are *underlined*. The *asterisk* marks the position of the conserved TDP-43 Phe residues with respect to the other RRM sequences.

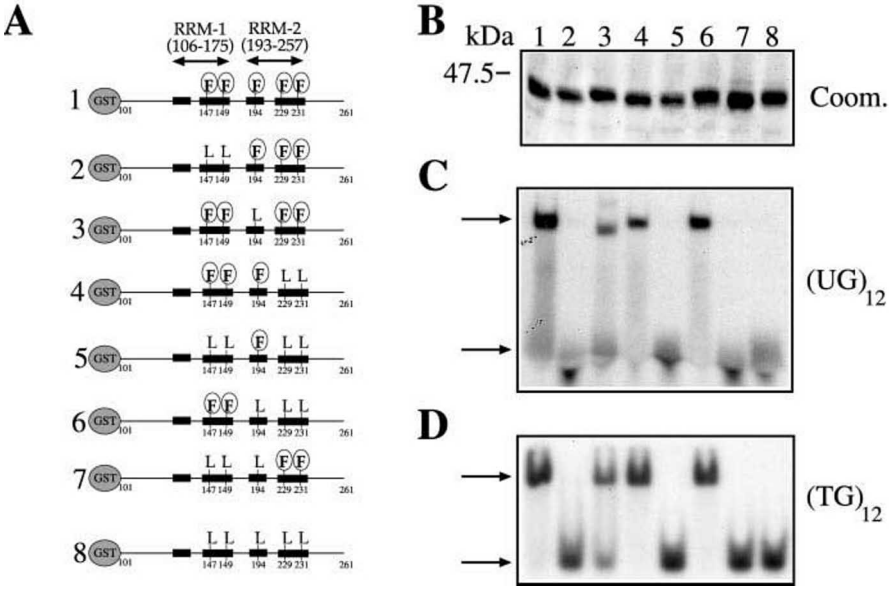

FIG. 6. **Importance of aromatic residues in TDP-43 RNP-1 and RNP-2 motifs for RNA/DNA binding efficiency.** *A,* panel of the Phe to Leu mutations (numbered 1–8) introduced in different RNP-1 and RNP-2 motifs of the GST-TDP43-(101–261) protein, either singly or in different combinations. *B,* a Coomassie Blue-stained SDS-PAGE gel of each mutant after purification with glutathione-Sepharose 4B beads according to standard protocols. *C* and *D,* EMSA of the effects of these mutations on the binding to labeled single-stranded $(UG)_{12}$ RNA (6 nM) and labeled $(TG)_{12}$ DNA (6 nM). Protein concentration for $(UG)_{12}$ RNA was 0.28 μM while for $(TG)_{12}$ was 0.14 μM. The numbering of each lane corresponds to the mutant used in each reaction mixture. The *arrows* on the *left* indicate the retarded protein-nucleic acid complexes (*upper arrow*) and the free nucleic acid (*lower arrow*).

These RNA binding characteristics of TDP-43 are also in good agreement with those of other similar RNA-binding proteins whose structure is known by crystallographic studies, such as hnRNP A1 (22), Sxl (23), and PABP (24) and whose RRMs share considerable homology with TDP-43 RRMs. In fact, the minimum length of single-stranded RNA bound by TDP-43 (a stretch of six UG repeats) is consistent with the length of single-stranded RNAs bound by similar proteins: $(a)_{11}$ in the case of PABP (24) and $ug(u)_7$ for Sxl (23). In this respect, therefore, TDP-43 acts very similarly to other well characterized proteins that also employ a two-RRM domain strategy to recognize RNA.

We have previously shown that an increase in TDP-43 cellular concentrations inhibits exon 9 splicing (1). This effect may not be necessarily mediated by simple binding competition with 3′-ss recognition factors but could also be linked to the still unknown cellular function of TDP-43. One alternative being that TDP-43 may bind to other cellular proteins that disrupt the recognition of exon 9 by the splicing machinery when positioned next to it. The search of the putative TDP-43 partners through protein similarity searches have been of limited use. In fact, they have shown that TDP-43 shares a high homology with a series of nuclear factors such as hnRNP D (38) or a mouse protein binding to CArG box motifs (39). However, the significance of these homologies is limited because they are predominantly localized in the central portion of TDP-43, which contains the highly conserved RRM motifs and the glycine-rich region, two elements that TDP-43 shares with many other RNA-binding proteins. Nonetheless, the cloning of homologous proteins in *Drosophila* (40) and *Caenorhabditis* (QZ0414) shows that TDP-43 is highly conserved. Interestingly, a comparison of its amino acid sequence with yeast proteins also shows a high identity (~30%) with two factors, HRP1 (Nuclear Polyadenylated RNA-binding protein, involved in pre-mRNA 3′-end cleavage and polyadenylation) (41, 42) and NSR1 (Nucleolar Protein involved in the processing of 20 S to 18 S rRNA) (43). In this respect, the homology with the NSR1 yeast protein is particularly interesting, because this protein has been described to bind (TG1–3) telomeric single-stranded repeated sequences (44). Alternatively, in the light of recent evidence, the ability of TDP-43 to bind single-stranded (TG)m repeats may also indicate its participation in the recombination process, and in this respect it is worthy to note that (CA/GT)n microsatellites repeats have been reported to affect homologous recombination in yeast meiosis (45) and GT repeats have also been associated with recombination frequency on human chromosome 22 (46). On the basis of these homologies it is then possible to speculate that the cellular function of TDP-43 may

concern some as yet unidentified facet of mRNA processing other than splicing. This hypothesis is also supported by the peculiarities of TDP-43 RNA/DNA binding properties when compared with classical RNA-binding proteins.

The results shown in Figs. 1 and 2 indicate that the RNA and DNA sequence binding specificities do not fully coincide. In fact, the TAR DNA sequence does not contain any (TG) repeats, and its only distinguishing feature are the two polypyrimidinic tracts, which at the RNA level are not substrates for TDP-43 binding. This is rather unusual if we consider that proteins such as hnRNP A1, which are also known to bind both single-stranded RNA and DNA, show very similar DNA/RNA sequence binding specificities (47, 48). Moreover, Fig. 3*C* shows the formation of multiple and distinct complexes with TG repeats as opposed to a single complex with UG repeats. Regarding the TDP-43 structure, deletion of TDP-43 RRM2 does not abolish RNA binding but results in the formation of a complex with altered mobility (Fig. 4*B*). However, a selective mutagenesis of the aromatic residues of RRM2 (Fig. 6*C, mutant 6*) results in the formation of a $(UG)_{12}$-TDP43 complex that has a mobility indistinguishable from the wild type. These results suggest that RRM1 supplies most of the requirements for specific RNA binding but also that elements present in RRM2 (aside from the aromatic residues) are needed for correct complex formation. This situation differs from what has been recently reported for UP1 where deletion of the second RRM motif did not affect the RNA binding properties of the protein (49) and suggests the presence of considerable interplay between RRM1 and RRM2 of TDP-43.

Acknowledgment—We thank Michela Zotti for skillful technical assistance.

REFERENCES

1. Buratti, E., Dork, T., Zuccato, E., Pagani, F., Romano, M., and Baralle, F. E. (2001) *EMBO J.* **20,** 1774–1784
2. Niksic, M., Romano, M., Buratti, E., Pagani, F., and Baralle, F. E. (1999) *Hum. Mol. Genet.* **8,** 2339–2349
3. Pagani, F., Buratti, E., Stuani, C., Romano, M., Zuccato, E., Niksic, M., Giglio, L., Faraguna, D., and Baralle, F. E. (2000) *J. Biol. Chem.* **275,** 21041–21047
4. Strong, T. V., Wilkinson, D. J., Mansoura, M. K., Devor, D. C., Henze, K., Yang, Y., Wilson, J. M., Cohn, J. A., Dawson, D. C., Frizzell, R. A., and Collins, F. S. (1993) *Hum. Mol. Genet.* **2,** 225–230
5. Delaney, S. J., Rich, D. P., Thomson, S. A., Hargrave, M. R., Lovelock, P. K., Welsh, M. J., and Wainwright, B. J. (1993) *Nat. Genet.* **4,** 426–431
6. Chillon, M., Casals, T., Mercier, B., Bassas, L., Lissens, W., Silber, S., Romey, M. C., Ruiz-Romero, J., Verlingue, C., Claustres, M., Nunes, V., Férec, C., and Estiril, X. (1995) *N. Engl. J. Med.* **332,** 1475–1480
7. Chu, C. S., Trapnell, B. C., Curristin, S., Cutting, G. R., and Crystal, R. G. (1993) *Nat. Genet.* **3,** 151–156
8. Cuppens, H., Lin, W., Jaspers, M., Costes, B., Teng, H., Vankeerberghen, A., Jorissen, M., Droogmans, G., Reynaert, I., Goossens, M., Nilius, B., and Cassiman, J. J. (1998) *J. Clin. Invest.* **101,** 487–496
9. Rave-Harel, N., Kerem, E., Nissim-Rafinia, M., Madjar, I., Goshen, R., Augarten, A., Rahat, A., Hurwitz, A., Darvasi, A., and Kerem, B. (1997) *Am. J. Hum. Genet.* **60,** 87–94
10. Shelley, C. S., and Baralle, F. E. (1987) *Nucleic Acids Res.* **15,** 3787–3799
11. Gabellini, N. (2001) *Eur. J. Biochem.* **268,** 1076–1083
12. Kazazian, H. H., Jr., and Moran, J. V. (1998) *Nat. Genet.* **19,** 19–24
13. Rozmahel, R., Heng, H. H., Duncan, A. M., Shi, X. M., Rommens, J. M., and Tsui, L. C. (1997) *Genomics* **45,** 554–561
14. Burd, C. G., and Dreyfuss, G. (1994) *Science* **265,** 615–621
15. Mattaj, I. W. (1993) *Cell* **73,** 837–840
16. Biamonti, G., and Riva, S. (1994) *FEBS Lett.* **340,** 1–8
17. Birney, E., Kumar, S., and Krainer, A. R. (1993) *Nucleic Acids Res.* **21,** 5803–5816
18. Nagai, K., Oubridge, C., Ito, N., Avis, J., and Evans, P. (1995) *Trends Biochem. Sci* **20,** 235–240
19. Ou, S. H., Wu, F., Harrich, D., Garcia-Martinez, L. F., and Gaynor, R. B. (1995) *J. Virol.* **69,** 3584–3596
20. Takahashi, N., Sasagawa, N., Suzuki, K., and Ishiura, S. (2000) *Biochem. Biophys. Res. Commun.* **277,** 518–523
21. Johnston, S. D., Lew, J. E., and Berman, J. (1999) *Mol. Cell. Biol.* **19,** 923–933
22. Shamoo, Y., Krueger, U., Rice, L. M., Williams, K. R., and Steitz, T. A. (1997) *Nat. Struct. Biol.* **4,** 215–222
23. Handa, N., Nureki, O., Kurimoto, K., Kim, I., Sakamoto, H., Shimura, Y., Muto, Y., and Yokoyama, S. (1999) *Nature* **398,** 579–585
24. Deo, R. C., Bonanno, J. B., Sonenberg, N., and Burley, S. K. (1999) *Cell* **98,** 835–845
25. Oubridge, C., Ito, N., Evans, P. R., Teo, C. H., and Nagai, K. (1994) *Nature* **372,** 432–438
26. Merrill, B. M., Stone, K. L., Cobianchi, F., Wilson, S. H., and Williams, K. R. (1988) *J. Biol. Chem.* **263,** 3307–3313
27. Jessen, T. H., Oubridge, C., Teo, C. H., Pritchard, C., and Nagai, K. (1991) *EMBO J.* **10,** 3447–3456
28. Hoffman, D. W., Query, C. C., Golden, B. L., White, S. W., and Keene, J. D. (1991) *Proc. Natl. Acad. Sci. U. S. A.* **88,** 2495–2499
29. Serin, G., Joseph, G., Ghisolfi, L., Bauzan, M., Erard, M., Amalric, F., and Bouvet, P. (1997) *J. Biol. Chem.* **272,** 13109–13116
30. Sutherland, G. R., and Richards, R. I. (1995) *Proc. Natl. Acad. Sci. U. S. A.* **92,** 3636–3641
31. International Human Genome Sequencing Consortium. (2001) *Nature* **409,** 860–920
32. Csink, A. K., and Henikoff, S. (1998) *Trends Genet.* **14,** 200–204
33. Redman, J. B., Fenwick, R. G., Jr., Fu, Y. H., Pizzuti, A., and Caskey, C. T. (1993) *J. Am. Med. Assoc.* **269,** 1960–1965
34. Yu, S., Pritchard, M., Kremer, E., Lynch, M., Nancarrow, J., Baker, E., Holman, K., Mulley, J. C., Warren, S. T., Schlessinger, D., Sutherland, G. R., and Richards, R. I. (1991) *Science* **252,** 1179–1181
35. Oberle, I., Rousseau, F., Heitz, D., Kretz, C., Devys, D., Hanauer, A., Boue, J., Bertheas, M. F., and Mandel, J. L. (1991) *Science* **252,** 1097–1102
36. The Huntington's Disease Collaborative Research Group. (1993) *Cell* **72,** 971–983
37. La Spada, A. R., Wilson, E. M., Lubahn, D. B., Harding, A. E., and Fischbeck, K. H. (1991) *Nature* **352,** 77–79
38. Kajita, Y., Nakayama, J., Aizawa, M., and Ishikawa, F. (1995) *J. Biol. Chem.* **270,** 22167–22175
39. Kamada, S., and Miwa, T. (1992) *Gene* (*Amst.*) **119,** 229–236
40. Lukacsovich, T., Asztalos, Z., Juni, N., Awano, W., and Yamamoto, D. (1999) *Genomics* **57,** 43–56
41. Wilson, S. M., Datar, K. V., Paddy, M. R., Swedlow, J. R., and Swanson, M. S. (1994) *J. Cell Biol.* **127,** 1173–1184
42. Anderson, J. T., Wilson, S. M., Datar, K. V., and Swanson, M. S. (1993) *Mol. Cell. Biol.* **13,** 2730–2741
43. Lee, W. C., Xue, Z. X., and Melese, T. (1991) *J. Cell Biol.* **113,** 1–12
44. Lin, J. J., and Zakian, V. A. (1994) *Nucleic Acids Res.* **22,** 4906–4913
45. Gendrel, C. G., Boulet, A., and Dutreix, M. (2000) *Genes Dev.* **14,** 1261–1268
46. Majewski, J., and Ott, J. (2000) *Genome Res.* **10,** 1108–1114
47. Buvoli, M., Cobianchi, F., Biamonti, G., and Riva, S. (1990) *Nucleic Acids Res.* **18,** 6595–6600
48. Abdul-Manan, N., and Williams, K. R. (1996) *Nucleic Acids Res.* **24,** 4063–4070
49. Dallaire, F., Dupuis, S., Fiset, S., and Chabot, B. (2000) *J. Biol. Chem.* **275,** 14509–14516