Acta Biochimica Polonica

Vol. 55 No. 1/2008, 1–8

on-line at: www.actabp.pl

Regular paper

Expression characteristics of triplet repeat-containing RNAs and triplet repeat-interacting proteins in human tissues

Anna J. Jasinska★, Piotr Kozlowski and Wlodzimierz J. Krzyzosiak✉

Laboratory of Cancer Genetics, Institute of Bioorganic Chemistry, Polish Academy of Sciences, Poznań, Poland

Received: 15 November, 2007; revised: 17 January, 2008; accepted: 29 January, 2008
available on-line: 30 January, 2008

Numerous human transcripts contain tandem repeats of trinucleotide motifs, the function of which remains unknown. In this study we used the available gene expression EST data to characterize the abundance of a large group of these transcripts in different tissues and determine the mRNAs which had the highest contribution to the observed levels of transcripts containing different types of the CNG repeats. A more extensive characteristics was performed for transcripts containing the CUG repeats, and those encoding the repeat-binding proteins. The scarcity of double-stranded CUG repeats as well as various proportions of the single-stranded and double-stranded CUG repeat-binding proteins were revealed in the studied transcriptomes. The observed correlated levels of transcripts containing single-stranded CUG repeats and of proteins binding single-stranded CUG repeats may imply that in addition to transcripts which only provide binding sites for these proteins there may be a substantial portion of the transcripts whose metabolism is directly regulated by such proteins. Our results showing a highly variable composition of triplet repeat-containing transcripts and their interacting proteins in different tissues may contribute to a better understanding of the mechanism of RNA-mediated pathogenesis in triplet repeat expansion diseases.

Keywords: RNA pathogenesis, EST, CAG-containing transcripts, CUG-containing transcripts

INTRODUCTION

Different types of microsatellites also known as simple sequence repeats (SSRs) or short tandem repeats (STRs) make a 3% contribution to the human genome. Importantly, repeated trimers and hexamers are more frequent in the coding regions of genes than in the intronic and intergenic sequences (Subramanian *et al.*, 2003a; 2003b; Piwowar *et al.*, 2006). The number of human genes containing at least four tandemly repeated trinucleotide sequences has been shown to exceed two thousand (Subramanian *et al.*, 2003a), i.e. about 10% of the estimated total number of human genes. The observed positive selection for triplet repeats in exons suggests some functions for these sequences which are, however, poorly known. The frequent length polymorphism of trinucleotide tracts makes them a rich source of human genetic variation required for evolutionary adaptation. The repeats may be involved in the regulation of gene expression at the various levels (Kashi *et al.*, 1997; Riley & Krieger, 2004), and their instability in different single genes causes a number of hereditary human disorders known as triplet repeat expansion diseases (TREDs).

The function of triplet repeats in transcripts is also barely known. When their occurrence in human mRNAs was analyzed it turned out that more than

✉Corresponding author: Wlodzimierz J. Krzyzosiak, Laboratory of Cancer Genetics, Institute of Bioorganic Chemistry, Polish Academy of Sciences, Noskowskiego 12/14, 61-704, Poznań, Poland; tel.: (48) 61 852 8503, fax.: (48) 61 852 0532; e-mail: wlodkrzy@ibch.poznan.pl

★Current address: Center for Neurobehavioral Genetics, University of California, Los Angeles, CA 90095, USA.

Abbreviations: CUG-BP, CUG-binding protein; DM, myotonic dystrophy; dsCUG, double-stranded CUG repeats; EST, expressed sequence tags; ssCUG, single-stranded CUG repeats; TREDs, triplet repeat expansion diseases; SSR, simple sequence repeat; STR, short tandem repeats; NCBI, National Center for Biotechnology Information.

600 mRNAs contained one or more tracts in which one of the 20 different trinucleotide sequences was repeated at least six times (Jasinska *et al.*, 2003). Among the repeated triplet motifs the CNG (N = A, G, T, C), AGG and ACC types strongly predominated. Most of the triplet repeats (67%) occurred in the open reading frame (24%) in the 5′ UTR which regulates mRNA translation, and 9% in the 3′ UTR which determines mRNA stability. What is noteworthy, GC-rich repeats predominated in the 5′ UTR, whereas AU-rich repeats were more frequent in 3′ UTR. In the open reading frame the CAG repeats were most abundant and they usually encoded polyglutamine tracts in proteins. Other CNG repeats were 2–3 times less frequent in translated sequences.

A common feature of the CNG repeats in transcripts is their tendency to form double-stranded (ds) hairpin structures if the repeat tracts are long enough (Napierala & Krzyzosiak, 1997; Sobczak *et al.*, 2003). The single-stranded (ss) structures and hairpins may serve as binding sites for dsRNA- and ssRNA-binding proteins, respectively (Timchenko *et al.*, 1996; Miller *et al.*, 2000; Fardaei *et al.*, 2001; Faustino & Cooper, 2003; Kino *et al.*, 2004). Proteins which bind ssCUG repeats belong to the CELF family (Ladd *et al.*, 2001), whereas proteins which bind dsCUG repeats belong to the muscleblind family (Napierala & Krzyzosiak, 1997; Miller *et al.*, 2000). The CELF proteins are known to regulate RNA splicing (Ladd *et al.*, 2001; Charlet *et al.*, 2002; Suzuki *et al.*, 2002; Ho *et al.*, 2004), editing (Anant *et al.*, 2001), and translation (Timchenko *et al.*, 1999; 2004). The muscleblind proteins also regulate splicing (Ho *et al.*, 2004) and play a yet unknown function in the cytoplasm. Proteins that bind to other types of the CNG repeats were only very preliminarily described (McLaughlin *et al.*, 1996; Tian *et al.*, 2000; Jin *et al.*, 2007; Sofola *et al.*, 2007). Normal functions of repeat-containing transcripts and repeat-binding proteins may require their balanced levels in cells (Jiang *et al.*, 2004). If their proportions are affected by transcripts containing expanded repeats, this may cause RNA-mediated pathogenesis in some TREDs (Miller *et al.*, 2000; Fardaei *et al.*, 2001; Mankodi *et al.*, 2003; Jiang *et al.*, 2004).

Several questions regarding the triplet repeat containing transcripts and their binding proteins have not been answered yet. These questions included: What is the abundance of such transcripts in different human tissues? Is their expression coordinated with the expression of their binding proteins? Which transcripts show the highest expression levels and could play the role of major acceptors of the CNG repeat-binding proteins? To answer these questions, we employed EST (expressed sequence tag) data from multiple normal human tissues available in the gene-centered database GeneCard displaying results of automatic mining of the UniGene (NCBI) resources (Safran *et al.*, 2003). Unlike sequence based hybridization methods which measure hybridization signal intensity of selected transcripts with a sequence-dependent efficiency, EST counts approximate absolute transcript levels allowing a direct comparison between mRNA products of different genes. The objects of our investigation presented here were human transcripts containing CNG repeat tracts. Previously, we used the Blast search to query all human transcripts deposited in the NCBI using sequences consisting of six pure CNG repeats, which led us to the identification of a multitude of CNG repeat-containing transcripts (Jasinska *et al.*, 2003). Here we make an attempt to investigate the role of the CNG repeats by examining their expression levels. In addition to transcript repeats, we estimate the amounts of the repeat-binding proteins. We made an assumption that RNA expression allows protein expression to be predicted and used the transcript levels as a proxy for the corresponding protein levels. Although there is an overall positive correlation between mRNA and protein levels, this simple correspondence between transcript and protein level needs to be interpreted with scrutiny, considering the posttranscriptional and posttranslational regulatory mechanisms. We employed the available EST expression data to examine the absolute tissue levels of CNG repeat-containing transcripts and transcripts coding for repeat-interacting proteins and here we discuss the resulting comprehensive picture of the expression levels of these transcripts in multiple human tissues.

MATERIALS AND METHODS

Data. Human EST counts (NCBI, UniGene) were taken from the GeneCards database which is accessible on http://www.genecards.org/index.shtml (Safran *et al.*, 2003). We used the EST tag counts from brain, skeletal muscle, liver, prostate, lung, kidney and pancreas tissues to determine the expression levels of individual repeat-containing transcripts as the number of EST clones assigned to this transcript in a given tissue.

Analyzed transcripts. The EST data were examined in the context of absolute expression levels in two groups of transcripts. The group of 67 transcripts harboring different types of CNG motifs repeated at least ten times was examined in five types of normal human tissues: brain, skeletal muscle, liver, prostate and lung, in which at least 45 000 ESTs were identified. The second group of the analyzed transcripts consisted of 56 mRNAs harboring at least six CUG repeats. Their expression was analyzed in six tissues (brain, prostate, liver, kidney, pancreas

and lung) for which the EST numbers exceeded 50 000. The CNG and CUG repeat-containing transcripts investigated here were identified previously (Jasinska *et al.*, 2003) and characterized with regard to the repeat length polymorphism (Rozanska *et al.*, 2007). Briefly, transcripts containing six and more trinucleotide repeats were extracted from the GenBank database by systematic BLAST search and the polymorphism of repeat sequence was characterized in 260 individuals from the general Polish population.

Additionally, seven transcripts coding for the known CUG repeat binding proteins (CUG-BPs): five *CELF* transcripts (*CUG-BP1*, *CUG-BP2*, *BRUNOL4*, *BRUNOL5* and *BRUNOL6*) and three *MBNLs* (*MBNL1*, *MBNL2* and *MBNL3*) were analyzed in these tissues.

Statistical methods. Chi squared test was used to compare distribution of transcripts containing different types of the CNG repeats in the five analyzed tissues. Expression levels of the CNG-binding proteins and CNG-containing transcripts were compared by linear correlation and R squared coefficient and *P*-value were calculated. All statistical analyses were performed using Statistica (StatSoft, Tulsa, OK, USA) or Prism v. 4.0 (GraphPad Software, San Diego, CA, USA).

RESULTS

Tissue expression profiles of the CNG repeat-containing transcripts vary considerably

The EST data obtained from the GeneCard browser (Safran *et al.*, 2003) was used to evaluate tissue levels of human transcripts containing the CNG motif repeated at least ten times. We searched for all human transcripts with this characteristics and the expression data was available for 67 such transcripts — mostly with CAG (39) and CGG (13) tracts — in at least one of the five tissues examined: brain, skeletal muscle, liver, prostate and lung, for which the relevant datasets were sufficiently large. We used the EST data to characterize the absolute levels of the transcripts in each tissue and differences between the transcripts. We observed that the number of individual transcripts differs considerably among the analyzed tissues (Supplementary Fig. 1 see: www.actabp.pl). The widest spectrum of different CNG-containing transcripts was observed in brain, which expresses 70% of the transcripts investigated (55). These included transcripts containing the CAG repeat (32), CUG (7), CGG (10) and CCG (6). The lowest number of different transcripts was observed in the liver (39). This follows the general tendency regarding the spectrum of gene expression in the brain and liver.

Interestingly, strong differences in expression are observed within the groups of transcripts bearing different repeat types, especially the CUG and CGG tracts. The *RPL14* alone contributes 83% and 54% to the total EST count from CUG-containing transcripts in the prostate and brain. *HSPC028* is another example of an individual transcript that makes up more than half of the CUG transcript pool (52%) in the lung. In the lung and prostate, a single gene — *CAPNS1* — gives rise to 37% of the ESTs that correspond to the CGG repeat-containing transcripts. Thus, it appears that individual highly abundant transcripts provide large quantities of the RNA repeats which may absorb a substantial portion of the repeat-binding proteins from their cellular pool.

CNG repeat-containing transcripts belong to moderate and low abundance classes

Given the expression levels of the CNG repeat containing transcripts determined from the EST data, we further assessed how these values correspond to absolute transcript levels in the analyzed tissues. We normalized each EST dataset to 150 000 tags and classified the expressed transcripts according to the number of their EST counts. To make this classification coherent with the expression levels of some reference transcripts, we also counted the ESTs corresponding to the *GAPD* and *ACTB* genes. Transcripts of these control genes are known to show high levels in most tissues and their normalized counts averaged over several tissues were 183 and 229, respectively. Accordingly, we classified the CNG repeat-containing transcripts represented by 16–50 ESTs to the moderate expression class, those represented by 8–15 ESTs corresponded to low expression, and those having 1–7 ESTs to the very low expression class. There were also numerous undetected transcripts in each tissue. The expression levels of transcripts containing CNG repeats were then between one and two orders of magnitude lower than those of the highly abundant reference transcripts (Fig. 1). The CNG-containing transcripts do not reach high expression levels but a substantial fraction of them (16–20%) is expressed at a moderate level in most of the examined tissues. Only in the liver, this abundance class is significantly underrepresented (4%). It is worth noting that moderately expressed transcripts, unlike the low abundance ones, are less prone to sampling errors affecting the number of detected transcripts and the differences in this abundance classes are more meaningful. Figure 1 shows that the different abundance classes of CNG transcripts are unevenly distributed among the tissues ($P = 0.04$; Chi squared test). Moreover, it is apparent

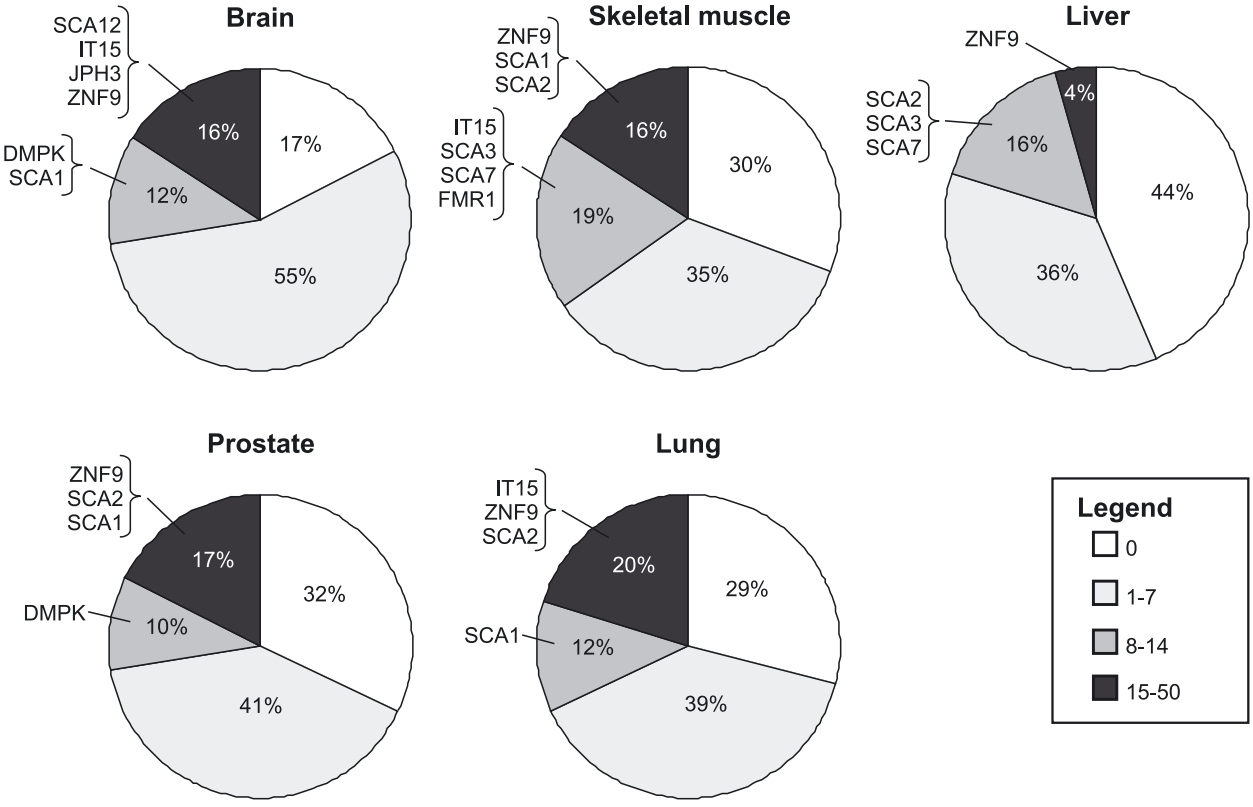

Figure 1. Expression classes of transcripts containing ten or more CNG repeats.
Transcript levels were determined from the EST data after normalization to 150 000. Analyzed transcripts were classified to different expression classes according to the number of corresponding EST tags: 0, 1–7, 8–15, > 15 ESTs. The TREDs genes classified as moderately or lowly expressed (> 8 normalized EST count) are shown in the diagram.

that the widest spectrum of transcripts observed in the brain results from the overrepresentation of the group of transcripts with a very low expression.

The set of the 67 analyzed CNG containing transcripts includes a total of 16 TREDs-related transcripts. As many as ten of mature TREDs transcripts (*SCA1*, *SCA2*, *SCA3*, *SCA7*, *SCA12*, *IT15*, *JPH3*, *AR*, *FMR1* and *DMPK*) show moderate expression levels in at least one of the investigated tissues (Fig. 1). For comparison with CNG repeat-containing TREDs transcripts we also included the *ZNF9* gene whose immature transcript contains CCUG repeats sharing some molecular properties with CNG repeats (Sobczak *et al.*, 2003). Notably, *ZNF9* is the most abundant TREDs-related transcript in as many as four out of the five tissues analyzed. Its highest expression is observed in the brain (45 ESTs) followed by liver (38 ESTs), prostate (28 ESTs) and skeletal muscle (20 ESTs). Only in the lung the *IT15* transcript is the most abundant (45 ESTs) followed by *ZNF9* (21 ESTs) and *SCA2* (17 ESTs). Three CAG repeat-containing tracts, *SCA12* (32 ESTs), *JPH3* (22 ESTs) and *IT15* (27 ESTs) are among the predominant TREDs-related transcripts in the brain. On the other hand, it is also worth noting that the *SCA8*, *FRDA*, *FMR2*, *SCA6* and *DRPLA* transcripts are either undetected or show very low expression in most of the tissues analyzed. Thus, the TREDs-related transcripts differ significantly in their expression both among the tissues and within each tissue, which parallels the trends observed in the whole group of the analyzed CNG repeat-containing transcripts.

Proportions between the analyzed CUG repeat-containing transcripts and transcripts encoding the repeat-binding proteins are different in different tissues

Of the different types of the CNG repeats, the knowledge regarding the functions of the CUG repeats in transcripts and of the proteins they bind is the most advanced (Miller *et al.*, 2000; Thornton *et al.*, 2003; Kino *et al.*, 2004). Accordingly, our further analysis was focused on the CUG-containing transcripts. We expanded the characterized above group to 56 transcripts harboring the CUG triplets repeated at least six times. To identify the putative major acceptors of the CUG repeat-binding proteins we analyzed the EST data from six human tissues (Supplementary Fig. 2 see: www.actabp.pl). The numbers of transcripts scoring more than 15 ESTs decreased in the order: the brain (15), lung (13), pancreas (13),

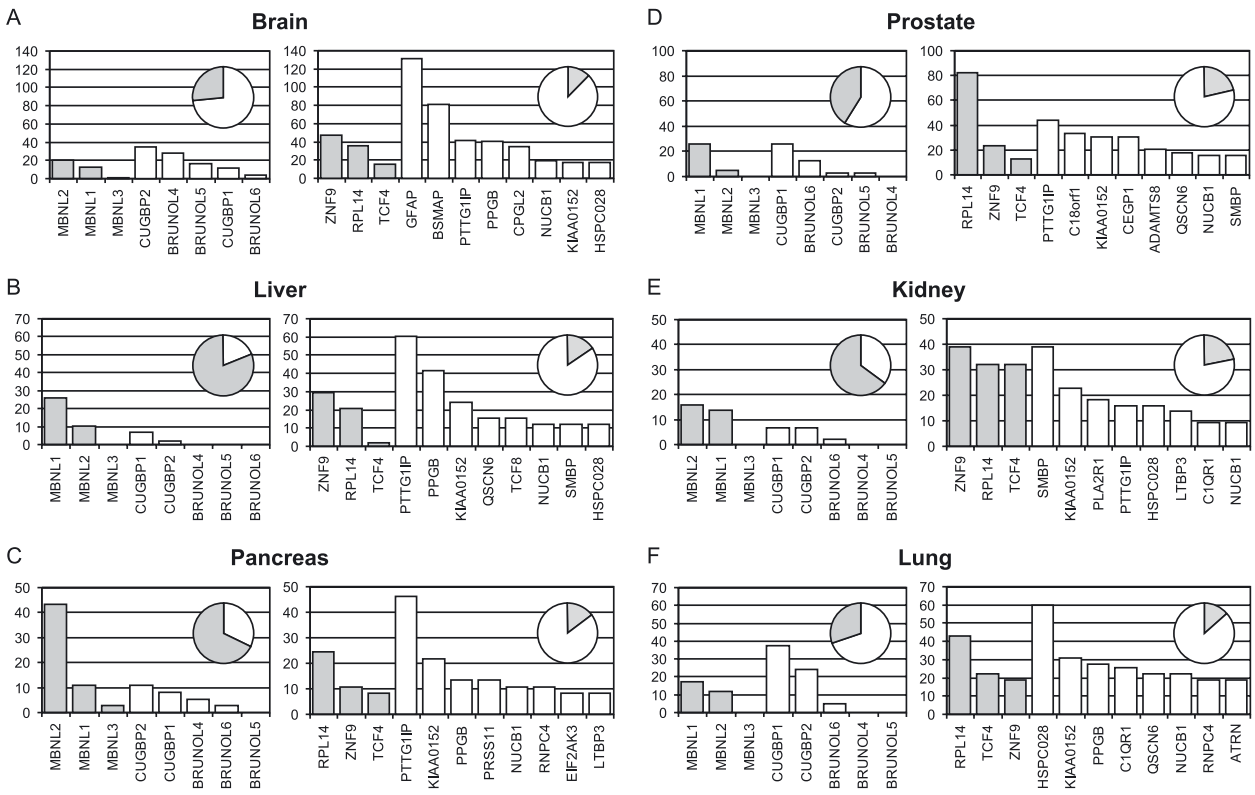

Figure 2. EST-based absolute levels of transcripts coding for eight different CUG-BPs and the most abundant of the analyzed CUG-containing transcripts.
The normalized EST data is shown for six normal human tissues: brain (A), liver (B), pancreas (C), prostate (D), kidney (E) and lung (F). Panel for each tissue consists of two bar diagrams with the EST count plotted on the vertical axis: the left-hand diagram shows the count for the *MBNLs* (gray bars) and *CELFs* (white bars), the right-hand diagram shows the count for transcripts that may contain short (white bars) and long (16 or more) (gray bars) CUG repeats. In the upper right-hand corner, the proportion of *ssCUG-BP* (white) and *dsCUG-BP* (gray) transcripts is shown (left diagram). A similar inset (right-hand diagram) shows the proportion of transcripts with long (gray) and short (white) CUG tracts. The abundance of each transcript is shown as the number of assigned ESTs in the dataset for a particular tissue after normalization to 150 000.

prostate (12), kidney (8) and liver (7). Importantly, several transcripts were represented by more than 50 ESTs in at least one tissue. These abundant transcripts included *GFAP* (131 ESTs) and *BSMAP* (82 ESTs) expressed in the brain and *PTTG11P* (60 ESTs) in liver. Some of these transcripts were frequent in all analyzed tissues (*PTTG11P*, *KIA0152*, *NUCB1*), whereas others in single tissues only, e.g. *GFAP*, *BSMAP* and *CPGL2* in the brain where they constituted more than 30% of all CUG repeat-containing transcripts.

To compare the absolute levels of the analyzed CUG repeat-containing transcripts and transcripts of their binding proteins, we gathered the relevant EST data for the latter transcripts corresponding to both the ssCUG and dsCUG repeat-binding proteins (Fig. 2). In the pancreas, liver and kidney the *MNBLs* transcripts predominate (*MBNL1* and *MBNL2* in particular) and the *CUG-BP1* and *CUG-BP2* transcripts show lower levels. Conversely, in the brain, lung and prostate transcripts of the ssCUG repeat-binding proteins are more prevalent (*CUG-BP1* and *CUG-BP2* in the lung and, in addition, *BRUNOLs* in the brain). Figure 2 shows proportions of the levels of the transcripts of both types of CUG-binding proteins and their most abundant targets. A general trend is that the levels of the ssCUG repeat-containing transcripts correlate well with the levels of the transcripts of the ssCUG repeat-binding proteins ($R^2 = 0.85$, $P < 0.01$), while no clear correlation can be detected for transcripts with dsCUG repeats and transcripts of their interacting proteins. Moreover, the combined EST data shows which of the CUG repeat containing transcripts are likely to determine the balance between the CELFs and MNBLs pool available in different tissues. For example, it became apparent that in the group of transcripts which are expressed in the brain tissue and are capable of binding CELFs, *GFAP*, *BSMAP* and *CPGL2* show high brain-specific abundance and may engage a substantial portion of the CELF proteins.

DISCUSSION

Not only different combinations of transcripts but also their highly variable levels specify numerous human transcriptomes characteristic for different cell types and tissues. Both the maturation of primary transcripts and the functions of mature RNAs are executed through their interactions with specific RNA-binding proteins, which very likely requires well balanced levels of the two types of interactors. The normal function of the CUG repeats in transcripts may also depend on their interactions with the specific repeat-binding proteins. As the structure of the CUG repeats in transcripts is not uniform and shorter repeats remain single-stranded whereas stable hairpins are formed by longer repeat tracts, also the proteins they bind belong, accordingly, to two distinct classes: the ssCUG-BP (CELFs) and dsCUG-BP (MBNLs). The RNA-binding properties of these proteins are beginning to be revealed. For example, it is known that MBNL1 binds the dsCUG repeats starting from (CUG)16 (Kino *et al.*, 2004) in a length-dependent manner (Miller *et al.*, 2000), and that CUG-BP1 binds even very short and irregular ssCUG repeats (Philips *et al.*, 1998; Charlet *et al.*, 2002; Suzuki *et al.*, 2002).

Our observations indicate that the proportions between the levels of the dsCUG-BPs and ssCUG-BPs transcripts vary among tissues, being strongly in favor of the CELFs in the brain and MBNLs in liver (Fig. 2). These proportions may suggest that the functions which require either the CELFs or MBNLs predominate in different tissues. Both classes of proteins are involved in the regulation of alternative splicing (Ladd *et al.*, 2001; Charlet *et al.*, 2002; Suzuki *et al.*, 2002; Ho *et al.*, 2004), and are known as antagonistic splicing regulators (Ho *et al.*, 2004). Thus, it may be anticipated that in tissues with different proportions of these proteins the transcripts they regulate may have opposite splicing patterns. So far, only a few transcripts regulated by the CELFs and MBNLs have been identified (Philips *et al.*, 1998; Savkur *et al.*, 2001; Charlet *et al.*, 2002; Suzuki *et al.*, 2002; Ho *et al.*, 2004; Paul *et al.*, 2006; Hino *et al.*, 2007). We previously identified and characterized with regard to repeat polymorphism numerous transcripts containing the CUG repeats which could participate in the CUG-BPs binding (Jasinska *et al.*, 2003; Rozanska *et al.*, 2007). In this study, by analyzing the levels of these transcripts we indicate which of them may belong to the major acceptors of the CUG repeat-binding proteins. Our present examination showing a clear correlation between the levels of transcripts containing ssCUG repeats and transcripts of ssCUG-binding proteins suggests that these balanced levels may be crucial for cellular functions. Whether the CUG-harboring transcripts may be somehow regulated by the CUG repeat-binding proteins or they only interfere with such regulatory processes, by lowering the pool of available CUG-BPs, remains to be established.

Interestingly, the great majority of the CUG repeat-containing transcripts harbor short single-stranded repeats which may be exclusively CELFs binders. Such transcripts strongly predominate in the transcriptomes of all six tissues analyzed. On the other hand, only some *DMPK*, *RPL14*, *ZNF9* and *TCF4* (*SEF2*) alleles, due to their repeat length, may give rise to transcripts capable of binding the MBNLs. Both *TCF4* and *ZNF9* have their repeat tracts located in introns and the nuclear levels of their primary transcripts may be high. *RPL14*, *ZNF9* and *TCF4*, which may bind the MBNLs, as well as *PTTG1P*, *NUCB1* and *KIA0152*, capable of binding CELFs, belong to transcripts showing considerable expression in a large spectrum of tissues. These transcripts may be among the major components of the putative regulatory networks controlled by the CUG repeat-binding proteins.

When the proportions between the CUG repeat-containing transcripts and their binding proteins are strongly imbalanced this may result in severe disorders. Myotonic dystrophy (Safran *et al.*, 2003) is an example of a human disease in which the mechanism of RNA-mediated pathogenesis is generally accepted (Ranum & Day, 2004). In this multi-system disease either the expansion of the CTG repeat located in the 3′ UTR of the *DMPK* gene (myotonic dystrophy type 1, DM1) or expansion of the CCTG repeat present in an intron of the *ZNF9* gene (myotonic dystrophy type 2, DM2) are established sources of pathogenesis. These mutations result in the formation of long stable hairpin structures formed by the repeats in transcripts (Napierala & Krzyzosiak, 1997; Michalowski *et al.*, 1999; Sobczak *et al.*, 2003). Both the long dsCUG and CCUG repeat containing transcripts recruit proteins of the MBNL family with which they co-localize in nuclear foci observed in DM cells (Fardaei *et al.*, 2002). In DM1 cells up-regulation of CUG-BP1 is also observed (Timchenko *et al.*, 2004). In agreement with the mechanism of RNA pathogenesis in DM the above effects compromise the normal functions of transcripts regulated by the CUG repeat-binding proteins. Our present study shows which transcripts, due to their relatively high expression levels, may belong to the major consumers of these proteins in the analyzed tissues (Fig. 2). For example, some alleles of the *RPL14* and *TCF4* transcripts have the MBNL-binding potential but the former is expressed at much higher levels in most of the analyzed tissues and thus may provide more binding sites for the protein. It is possible that the normal functions of these transcripts might be compromised by the expanded repeats in the *DMPK*

and *ZNF9* transcripts. In this context it should also be noted that the *ZNF9* transcript is expressed at higher levels than *DMPK* not only in skeletal muscles (Mankodi *et al.*, 2003) but also in five other tissues analyzed here. Moreover, its expanded repeat tracts may be longer than the lengths of the CUG repeats in *DMPK* transcript (Liquori *et al.*, 2001). Thus, looking from this perspective *ZNF9* should be a more potent effector of MBNL sequestration. On the other hand, it seems that functions of a much higher number of transcripts may be influenced by the CUG-BP1 up-regulation in DM1 cells.

The results of this study also give some new clues in an attempt to answer the old question: why pathology develops in some cell types and tissues only, while its molecular triggers, in the form of mutant transcripts or proteins, show a more widespread expression? It is apparent that the expanded repeat containing transcripts or aberrant polyglutamine-containing proteins act in the background of the RNA–protein and protein–protein interactomes which have different compositions in various cell types and tissues including their different physiological conditions and developmental stages. Some of these different backgrounds may be prohibitive for the pathology to fully develop. They may allow, however, some of the disease hallmarks to appear. The recent findings of nuclear and cytoplasmic inclusions present in neurons (Jiang *et al.*, 2004) and fibroblasts (Fardaei *et al.*, 2001) of myotonic dystrophy type 1 (DM1) patients are among the facts supporting the above scenario.

It is not unlikely that, similar to the CUG repeats and their binding proteins, also other repeats of the CNG family may be involved in normal regulatory processes in cells and trigger pathogenesis when expanded. In support of this notion are the deleterious effects caused by the expanded CGG repeats in the *FMR1* transcript in some "premutation" carriers in the corresponding gene (Hagerman & Hagerman, 2004). Recently two groups presented data indicating that altered RNA–protein interactions are involved in pathogenesis of FAXTAS (Jin *et al.*, 2007; Sofola *et al.*, 2007; Swanson & Orr, 2007). These repeats form hairpin structures whose lengths correspond well to the pathogenic thresholds (Napierala *et al.*, 2005). Moreover, also the lengths of the hairpin structures formed by the uninterrupted CAG repeats in the *SCA1* and *SCA2* transcripts correlate well with the pathogenesis of these spinocerebellar ataxias (Sobczak & Krzyzosiak, 2004; 2005).

Acknowledgements

This work was supported by funding under the Sixth Research Framework Programme of the European Union, Project RIGHT (LSHB-CT-2004-005276) and by the Ministry of Science and Higher Education, grants: PBZ-KBN-124/P05/2004 and PBZ-MNiI-2/1/2005.

REFERENCES

Anant S, Henderson JO, Mukhopadhyay D, Navaratnam N, Kennedy S, Min J, Davidson NO (2001) Novel role for RNA-binding protein CUGBP2 in mammalian RNA editing. CUGBP2 modulates C to U editing of apolipoprotein B mRNA by interacting with apobec-1 and ACF, the apobec-1 complementation factor. *J Biol Chem* **276**: 47338–47351.

Charlet BN, Savkur RS, Singh G, Philips AV, Grice EA, Cooper TA (2002) Loss of the muscle-specific chloride channel in type 1 myotonic dystrophy due to misregulated alternative splicing. *Mol Cell* **10**: 45–53.

Fardaei M, Larkin K, Brook JD, Hamshere MG (2001) *In vivo* co-localisation of MBNL protein with DMPK expanded-repeat transcripts. *Nucleic Acids Res* **29**: 2766–2771.

Fardaei M, Rogers MT, Thorpe HM, Larkin K, Hamshere MG, Harper PS, Brook JD (2002) Three proteins, MBNL, MBLL and MBXL, co-localize *in vivo* with nuclear foci of expanded-repeat transcripts in DM1 and DM2 cells. *Hum Mol Genet* **11**: 805–814.

Faustino NA, Cooper TA (2003) Pre-mRNA splicing and human disease. *Genes Dev* **17**: 419–437.

Hagerman PJ, Hagerman RJ (2004) The fragile-X premutation: a maturing perspective. *Am J Hum Genet* **74**: 805–816.

Hino S, Kondo S, Sekiya H, Saito A, Kanemoto S, Murakami T, Chihara K, Aoki Y, Nakamori M, Takahashi MP, Imaizumi K (2007) Molecular mechanisms responsible for aberrant splicing of SERCA1 in myotonic dystrophy type 1. *Hum Mol Genet* **16**: 2834–2843.

Ho TH, Charlet BN, Poulos MG, Singh G, Swanson MS, Cooper TA (2004) Muscleblind proteins regulate alternative splicing. *EMBO J* **23**: 3103–3112.

Jasinska A, Michlewski G, de Mezer M, Sobczak K, Kozlowski P, Napierala M, Krzyzosiak WJ (2003) Structures of trinucleotide repeats in human transcripts and their functional implications. *Nucleic Acids Res* **31**: 5463–5468.

Jiang H, Mankodi A, Swanson MS, Moxley RT, Thornton CA (2004) Myotonic dystrophy type 1 is associated with nuclear foci of mutant RNA, sequestration of muscleblind proteins and deregulated alternative splicing in neurons. *Hum Mol Genet* **13**: 3079–3088.

Jin P, Duan R, Qurashi A, Qin Y, Tian D, Rosser TC, Liu H, Feng Y, Warren ST (2007) Pur alpha binds to rCGG repeats and modulates repeat-mediated neurodegeneration in a *Drosophila* model of fragile X tremor/ataxia syndrome. *Neuron* **55**: 556–564.

Kashi Y, King D, Soller M (1997) Simple sequence repeats as a source of quantitative genetic variation. *Trends Genet* **13**: 74–78.

Kino Y, Mori D, Oma Y, Takeshita Y, Sasagawa N, Ishiura S (2004) Muscleblind protein, MBNL1/EXP, binds specifically to CHHG repeats. *Hum Mol Genet* **13**: 495–507.

Ladd AN, Charlet N, Cooper TA (2001) The CELF family of RNA binding proteins is implicated in cell-specific and developmentally regulated alternative splicing. *Mol Cell Biol* **21**: 1285–1296.

Liquori CL, Ricker K, Moseley ML, Jacobsen JF, Kress W, Naylor SL, Day JW, Ranum LP (2001) Myotonic dys-

trophy type 2 caused by a CCTG expansion in intron 1 of ZNF9. *Science* **293**: 864–867.

Mankodi A, Teng-Umnuay P, Krym M, Henderson D, Swanson M, Thornton CA (2003) Ribonuclear inclusions in skeletal muscle in myotonic dystrophy types 1 and 2. *Ann Neurol* **54**: 760–768.

McLaughlin BA, Spencer C, Eberwine J (1996) CAG trinucleotide RNA repeats interact with RNA-binding proteins. *Am J Hum Genet* **59**: 561–569.

Michalowski S, Miller JW, Urbinati CR, Paliouras M, Swanson MS, Griffith J (1999) Visualization of double-stranded RNAs from the myotonic dystrophy protein kinase gene and interactions with CUG-binding protein. *Nucleic Acids Res* **27**: 3534–3542.

Miller JW, Urbinati CR, Teng-Umnuay P, Stenberg MG, Byrne BJ, Thornton CA, Swanson MS (2000) Recruitment of human muscleblind proteins to $(CUG)_{(n)}$ expansions associated with myotonic dystrophy. *EMBO J* **19**: 4439–4448.

Napierala M, Krzyzosiak WJ (1997) CUG repeats present in myotonin kinase RNA form metastable "slippery" hairpins. *J Biol Chem* **272**: 31079–31085.

Napierala M, Michalowski D, de Mezer M, Krzyzosiak WJ (2005) Facile FMR1 mRNA structure regulation by interruptions in CGG repeats. *Nucleic Acids Res* **33**: 451–463.

Paul S, Dansithong W, Kim D, Rossi J, Webster NJ, Comai L, Reddy S (2006) Interaction of muscleblind, CUG-BP1 and hnRNP H proteins in DM1-associated aberrant IR splicing. *EMBO J* **25**: 4271–4283.

Philips AV, Timchenko LT, Cooper TA (1998) Disruption of splicing regulated by a CUG-binding protein in myotonic dystrophy. *Science* **280**: 737–741.

Piwowar M, Meus J, Piwowar P, Wisniowski Z, Stefaniak J, Roterman I (2006) Tandemly repeated trinucleotides — comparative analysis. *Acta Biochim Polon* **53**: 279–287.

Ranum LP, Day JW (2004) Pathogenic RNA repeats: an expanding role in genetic disease. *Trends Genet* **20**: 506–512.

Riley DE, Krieger JN (2004) Simple repeat replacements support similar functions of distinct repeats in interspecies mRNA homologs. *Gene* **328**: 17–24.

Rozanska M, Sobczak K, Jasinska A, Napierala M, Kaczynska D, Czerny A, Koziel M, Kozlowski P, Olejniczak M, Krzyzosiak WJ (2007) CAG and CTG repeat polymorphism in exons of human genes shows distinct features at the expandable loci. *Hum Mutat* **28**: 451–458.

Safran M, Chalifa-Caspi V, Shmueli O, Olender T, Lapidot M, Rosen N, Shmoish M, Peter Y, Glusman G, Feldmesser E *et al.* (2003) Human Gene-Centric Databases at the Weizmann Institute of Science: GeneCards, UDB, CroW 21 and HORDE. *Nucleic Acids Res* **31**: 142–146.

Savkur RS, Philips AV, Cooper TA (2001) Aberrant regulation of insulin receptor alternative splicing is associated with insulin resistance in myotonic dystrophy. *Nat Genet* **29**: 40–47.

Sobczak K, Krzyzosiak WJ (2004) Imperfect CAG repeats form diverse structures in SCA1 transcripts. *J Biol Chem* **279**: 41563–41572.

Sobczak K, Krzyzosiak WJ (2005) CAG repeats containing CAA interruptions form branched hairpin structures in spinocerebellar ataxia type 2 transcripts. *J Biol Chem* **280**: 3898–3910.

Sobczak K, de Mezer M, Michlewski G, Krol J, Krzyzosiak WJ (2003) RNA structure of trinucleotide repeats associated with human neurological diseases. *Nucleic Acids Res* **31**: 5469–5482.

Sofola OA, Jin P, Qin Y, Duan R, Liu H, de Haro M, Nelson DL, Botas J (2007) RNA-binding proteins hnRNP A2/B1 and CUGBP1 suppress fragile X CGG premutation repeat-induced neurodegeneration in a *Drosophila* model of FXTAS. *Neuron* **55**: 565–571.

Subramanian S, Madgula VM, George R, Mishra RK, Pandit MW, Kumar CS, Singh L (2003a) Triplet repeats in human genome: distribution and their association with genes and other genomic regions. *Bioinformatics* **19**: 549–552.

Subramanian S, Mishra RK, Singh L (2003b) Genome-wide analysis of microsatellite repeats in humans: their abundance and density in specific genomic regions. *Genome Biol* **4**: R13.

Suzuki H, Jin Y, Otani H, Yasuda K, Inoue K (2002) Regulation of alternative splicing of alpha-actinin transcript by Bruno-like proteins. *Genes Cells* **7**: 133–141.

Swanson MS, Orr HT (2007) Fragile X tremor/ataxia syndrome: blame the messenger! *Neuron* **55**: 535–537.

Thornton C, Swanson M, Cooper T (2003) The RNA-mediated disease process in myotonic dystrophy. In *Genetic Instabilities and Neurological Diseases*. Wells R, Ashizawa T, eds, pp 37–54. Academic Press.

Tian B, White RJ, Xia T, Welle S, Turner DH, Mathews MB, Thornton CA (2000) Expanded CUG repeat RNAs form hairpins that activate the double-stranded RNA-dependent protein kinase PKR. *RNA* **6**: 79–87.

Timchenko LT, Miller JW, Timchenko NA, DeVore DR, Datar KV, Lin L, Roberts R, Caskey CT, Swanson MS (1996) Identification of a (CUG)n triplet repeat RNA-binding protein and its expression in myotonic dystrophy. *Nucleic Acids Res* **24**: 4407–4414.

Timchenko NA, Welm AL, Lu X, Timchenko LT (1999) CUG repeat binding protein (CUGBP1) interacts with the 5′ region of C/EBPbeta mRNA and regulates translation of C/EBPbeta isoforms. *Nucleic Acids Res* **27**: 4517–4525.

Timchenko NA, Patel R, Iakova P, Cai ZJ, Quan L, Timchenko LT (2004) Overexpression of CUG triplet repeat-binding protein, CUGBP1, in mice inhibits myogenesis. *J Biol Chem* **279**: 13129–13139.