6260–6268 *Nucleic Acids Research, 2008, Vol. 36, No. 19* *Published online 2 October 2008*
doi:10.1093/nar/gkn511

G-quadruplexes: the beginning and end of UTRs

Julian Leon Huppert[1,*], Anthony Bugaut[2], Sunita Kumari[2] and Shankar Balasubramanian[2]

[1]Cavendish Laboratory, University of Cambridge, JJ Thompson Ave, Cambridge CB3 0HE and
[2]University Chemical Laboratory, University of Cambridge, Lensfield Road, Cambridge CB2 1EW, UK

Received April 18, 2008; Revised July 20, 2008; Accepted July 25, 2008

ABSTRACT

Molecular mechanisms that regulate gene expression can occur either before or after transcription. The information for post-transcriptional regulation can lie within the sequence or structure of the RNA transcript and it has been proposed that G-quadruplex nucleic acid sequence motifs may regulate translation as well as transcription. Here, we have explored the incidence of G-quadruplex motifs in and around the untranslated regions (UTRs) of mRNA. We observed a significant strand asymmetry, consistent with a general depletion of G-quadruplex-forming RNA. We also observed a positional bias in two distinct regions, each suggestive of a specific function. We observed an excess of G-quadruplex motifs towards the 5′-ends of 5′-UTRs, supportive of a hypothesis linking 5′-UTR RNA G-quadruplexes to translational control. We then analysed the vicinity of 3′-UTRs and observed an over-representation of G-quadruplex motifs immediately after the 3′-end of genes, especially in those cases where another gene is in close proximity, suggesting that G-quadruplexes may be involved in the termination of gene transcription.

INTRODUCTION

There are numerous mechanisms that regulate gene expression either at the DNA level or at the RNA level. Mechanisms for post-transcriptional regulation include the control of mRNA processing, nucleocytoplasmic transport, cellular and subcellular localization, translation efficiency and stability. Several studies have demonstrated that the genetic information required for post-transcriptional control is located mainly in the 5′- and 3′-untranslated regions (UTRs) of mRNA, and may involve both the primary sequence and secondary structure of non-protein-coding elements (1). Control via primary sequence recognition is exemplified by the action of ~22 nt single-stranded *trans*-acting regulatory elements, called microRNAs, that target mRNA sites, generally in their 3′-UTR, leading to translation repression (2). Secondary structures formed in 5′- and 3′-UTRs can also serve as regulatory elements by acting as target sites for RNA-binding factors such as proteins (3) or small molecule metabolites (4), or by interacting directly with the translation machinery (5–7).

There is scope for non-canonical nucleic acid structures to form in RNA transcripts. Certain guanine-rich nucleic acid sequences are predisposed to adopting four-stranded structures known as G-quadruplexes (8) that comprise stacks of hydrogen bonded G-tetrads, each containing four guanines. There is evidence that G-quadruplexes can form in the DNA at telomeres (10) under the control of telomere-binding proteins (9,10). Owing to the functional relationship between telomere maintenance, cell proliferation, and cancer, the telomeric DNA G-quadruplex is under consideration as a potential molecular target for anticancer therapeutics (11). It has also been proposed that DNA G-quadruplex motifs found within gene promoters may be involved in controlling gene expression at the transcriptional level (12). There is some *in vitro* experimental evidence for the promoter-quadruplex hypothesis from chemical biology studies on proto-oncogenes that including c-myc (13), k-ras (14) and c-kit (15). Furthermore, genome-wide computational analysis has revealed that putative quadruplex-forming sequences (putative quadruplex sequences, PQS) are prevalent in the human genome (16,17), and that there is a significant enrichment in gene promoter regions relative to the rest of the genome, with almost half of all protein-coding genes found to have putative quadruplex-forming motifs in their promoters (18). This has also been found to be the case for a variety of warm-blooded animals and the presence of putative G-quadruplex motifs in the first intron of genes has recently been studied (19).

While significant attention has hitherto been focused on DNA G-quadruplexes and their potential role in biology, certain G-rich RNA sequences can also fold into quadruplexes (20,21). Indeed, a few cases have been reported where intramolecular G-quadruplex formation within mRNAs has been proposed to be associated with function. For example, G-quadruplex formation in the 3′-UTR of insulin-like growth factor IGF-II mRNA was shown to

*To whom correspondence should be addressed. Tel: +44 1223 337 256; Fax: +44 1223 337 000; Email: jlh29@cam.ac.uk
Correspondence may also be addressed to Shankar Balasubramanian. Tel: +44 1223 336 347; Fax: +44 1223 336 913; Email: sbl0031@cam.ac.uk

occur downstream of an endonucleolytic cleavage site (22), the fragile X mental retardation protein (FMRP) has been shown to bind a G-quadruplex within the coding region of the corresponding mRNA (23) and an intramolecular RNA G-quadruplex motif has been found within the fibroblast growth factor (FGF-2) internal ribosome entry site (24). A cytoplasmic exoribonuclease, mXRN1p, has been shown to exhibit a substrate preference for G-quadruplex RNA (25). We recently discovered a conserved, intramolecular G-quadruplex motif within the 5′-UTR of the gene transcript of the human *NRAS* proto-oncogene, and we have demonstrated that this RNA G-quadruplex inhibits translation (26). Computational analysis revealed that there are 2922 genes containing 5′-UTR RNA G-quadruplex elements in the human genome. Herein, we report on a detailed computational study of putative quadruplex-forming sequences associated with the 5′- and 3′-UTRs of human mRNAs. The outcomes of this study have provided the basis for proposals as to how such motifs may be involved in the regulation of gene expression.

METHODS

Sequence data and gene descriptions were extracted from Ensembl with biomart, using build 36 of the human genome sequence throughout. Sequences analysed were known transcripts of known protein-coding genes, using the Ensembl definitions and flags throughout. We have considered every transcript individually, unless they had identical UTR sequences; thus there are in some cases more than one UTR per gene. We investigated the key conclusions of this work using only genes with precisely one transcript; the results were broadly the same, and are shown in supplementary material. Where gene-level analyses are reported, a gene was considered to have a UTR PQS if any of its transcripts did so. PQS were identified using the program *quadparser* (16), which is available online at http://www.quadruplex.org/?view=quadparser. Briefly, we used the default parameters for this program, which searches for sequences of the form $G_{3+}N_{1-7}G_{3+}N_{1-7}G_{3+}N_{1-7}G_{3+}$ on either strand of the sequence given. Other analyses were performed using custom-written perl scripts. Statistical analyses were performed using chi-squared tests or as otherwise appropriate. Full details are available in Supplementary material.

RESULTS AND DISCUSSION

Computational approaches can provide insights into the potential roles of sequence motifs that cluster in specific structural regions of the genome. We have previously used computational analysis of genomic data to suggest a functional role for G-quadruplex sequence motifs within gene promoters (18). Here, we have mapped the location of G-quadruplex motifs with high resolution in and around the 5′- and 3′-UTRs of human protein-coding genes.

Incidence of G-quadruplex motifs in 5′-UTRs and in 3′-UTRs

Using build 36 of the human genome sequence, we extracted unique 5′- and 3′-UTRs corresponding to all known protein-coding transcripts post-splicing. This yielded 21 658 unique genes, with a total of 32 985 annotated 5′-UTRs, and 32 818 3′-UTRs. The 5′-UTRs were in general significantly shorter than the 3′-UTRs, with a mean length of 243 bases, compared with 899 bases for the 3′-UTRs. The median lengths were 120 and 494 bases, respectively, showing that the distribution of UTR lengths comprised many relatively short UTRs and a tail of long UTRs, with the longest 5′-UTR being 24 kb, and the longest 3′-UTR 14 kb. From these data, using our search program *quadparser* (16), we investigated whether these UTRs contain PQS, on either of the two strands present in the cDNA sequence (Table 1). To distinguish between the two strands, we have referred to motifs identified in the *coding* strand as G-PQS (as the transcribed RNA sequence would contain a G-rich sequence capable of forming a G-quadruplex), and motifs identified in the *template* strand as C-PQS (as the transcribed RNA sequence would be C-rich and not form a G-quadruplex) (Figure 1).

Table 1. Summary of results for the association of 5′- and 3′-UTRs with PQS

	5′ UTRs	3′ UTRs
No. UTRs	32 985	32 818
Average length	243 bases	899 bases
No. UTRs with PQS (%)	4141 (12.6%)	5041 (15.3%)
With G-PQS	2034 (6.2%)	2740 (8.3%)
With C-PQS	2525 (7.7%)	3252 (9.9%)
Ratio C/G	1.24	1.19
No. G-PQS	2334	3530
No. C-PQS	3070	4526
G-PQS density	0.291/kb	0.120/kb
C-PQS density	0.382/kb	0.153/kb
Ratio C/G	1.32	1.28
Transcriptome G-PQS	0.077/kbase	
Transcriptome C-PQS	0.077/kbase	
Whole genome G-PQS	0.057/kbase	
Whole genome C-PQS	0.057/kbase	

The number of UTRs with G-PQS and the number with C-PQS do not sum to the total number with PQS, as some UTRs contain both a G-PQS and a C-PQS.

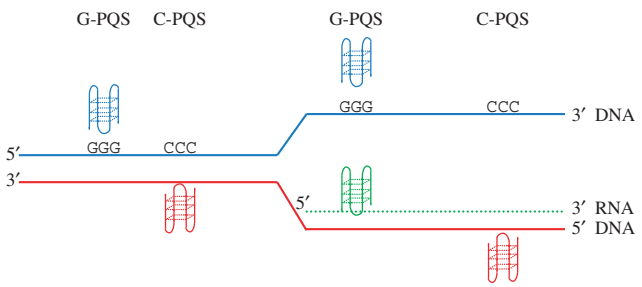

Figure 1. Schematic of DNA and transcribed RNA. Where the DNA sequence in the coding strand (blue) is G-rich (shown as GGG) a DNA G-quadruplex could form in that strand, and is here referred to as a G-PQS. A C-rich region in the coding strand is shown equivalently as CCC, and would allow a G-quadruplex to form on the template strand (red) and is here referred to as a C-PQS. After transcription, G-PQS, but not C-PQS, also results in the formation of a G-quadruplex in the mRNA (green).

Of the 32 985 5′-UTRs, 4141 (12.6%) exhibited one or more PQS on one of the two strands. However, the two strands were not equivalent, with only 2034 (6.2%) having a G-PQS, whereas 2525 (7.7%) were associated with a C-PQS. We also calculated the overall densities of PQS in 5′-UTRs to be 0.291 G-PQS/kb, and 0.382 C-PQS/kb. This shows a significantly greater proportion of C-PQS than G-PQS by a factor of 1.31 ($P = 3 \times 10^{-29}$). Of the 32 818 3′-UTRs, 5041 (15.3%) exhibited one or more PQS on one of the two strands. The proportion of 3′-UTR with PQS is therefore higher, but this could be easily explained (indeed, overcompensated for) by the observation that 3′-UTRs are in general much larger that 5′-UTRs, by a factor of about 4. On considering the two strands separately, we found that in the 3′-UTR, 2740 (8.3%) were associated with G-PQS, and 3252 (9.9%) with C-PQS. However, the densities of the two motifs are significantly lower ($P = 1 \times 10^{-53}$) in the 3′-UTR as compared with the 5′-UTR, at 0.120 G-PQS/kb and 0.153 C-PQS/kb. This gives a strand asymmetry ratio of 1.28 ($P = 2 \times 10^{-24}$) for 3′-UTR PQS, broadly the same as for PQS in the 5′-UTRs. These results are summarized in Table 1.

The occurrence of G-quadruplex motifs is strongly affected by base composition. Therefore, we investigated whether the asymmetry observed could be accounted for by this factor. In the 5′-UTR, there are more G bases than C bases (29.8% against 28.8%), which would suggest that more G-PQS would be expected to be present than C-PQS, in contradiction to the observed result. In the 3′-UTR, both bases are present in almost exactly equal proportions, at 21.7 and 21.6%, respectively. This could account for the difference in PQS density between the 5′- and 3′-UTRs, but not the excess presence of C-PQS.

To further study this effect, we generated simulated UTRs. To make these simulates, we studied every UTR, and counted the frequency with which each base was found at every position from the 5′-end. For each position, we counted the number of times the real UTRs terminated at that position. We then generated simulates in which the base frequencies at every position reflected the natural base frequencies at that position, and with UTR termination occurring at each position depending on the natural termination probability. One hundred replicates each consisting of 100 000 5′- and 3′-UTRs were generated, and searched for G-quadruplex motifs as above. As expected from the base frequencies, G-PQS were more frequently observed than C-PQS. For the 5′-UTRs, the simulates gave a ratio of C-PQS/G-PQS of 0.67 ± 0.09, compared to an observed ratio of 1.31 (significant, $P = 5 \times 10^{-14}$ based on the normal distribution curve). For the 3′-UTRs the simulated asymmetry was lower, at 0.96 ± 0.10, compared to the observed ratio of 1.28 (significant, $P = 4 \times 10^{-4}$ based on the normal distribution curve). Thus, base composition alone cannot account for the observed effects.

For comparison, we also examined the G-quadruplex densities of the entire transcriptome (defined as the total transcript of known human genes, including introns, according to Ensembl) as well as the whole genome [as described previously (16)]. Both of these showed virtually no strand asymmetry. In the transcriptome as a whole (1.1 Gb) we found 86 472 G-PQS and 86 038 C-PQS. This gives a transcriptome PQS density of 0.077 PQS/kb for each strand, for a total of 0.153 PQS/kb overall, slightly higher than that for the whole genome, which had an overall density of 0.115 PQS/kb, also split equally between the two strands.

Next, we analysed the Gene Ontology (GO) codes associated with all of these genes, to see if they corresponded to any over- or under-represented categories of genes (Tables 2 and 3). We used the same methodology reported in our previous analysis of gene promoters (18), which compared the number of genes in a particular GO category with G-PQS in a region to the total number of genes in that GO category, which varies significantly. We found that for the 5′-UTR G-PQS, 22 GO categories were significant at our required threshold of $P < 3.9 \times 10^{-6}$ (using the conservative Bonferroni correction to a P-value of 0.05, considering two tests on each of 6447 GO codes). For the 3′-UTR G-PQS, 28 GO categories were significant at the same level. All of these categories, together with the associated P-values, are listed in Tables 2 and 3. Twelve of these GO categories were the same as for the 5′-end, again showing a relationship between the two UTRs. Interestingly, some of the GO categories shown to be unlikely to have promoter PQS (18) were also found to be very unlikely to have G-PQS in either their 5′- or 3′-UTRs, such as the genes involved in olfaction and immune response.

Association of G-quadruplex motifs in *both* 3′- and 5′-UTRs

We considered whether genes with G-quadruplexes in their 5′-UTRs were also more likely to have them in their 3′-UTRs, or whether these were independent properties. Here, we have restricted our analysis to only include G-PQS sequences, as there is some evidence to support the hypothesis that these motifs may be associated with function in RNA, although the C-PQS could potentially form DNA G-quadruplexes on the template strand or a C-rich RNA secondary structure called the i-motif (27,28).

We found that 314 genes had G-PQS motifs in both the 5′- and 3′-UTRs, out of a total of 1665 genes with G-PQS in the 5′-UTR and 2154 with G-PQS in the 3′-UTR, from an overall total of 21 658 genes (Figure 2). The proportion of genes with 5′-UTR G-PQS that also have 3′-UTR PQS was 18.9% (314/1665) as compared to the proportion of all genes with 3′-UTR PQS, which was 9.9% (2154/21 658). If these two sets of motifs were entirely independent, the two proportions would necessarily be identical. Thus, there is clearly a significant association ($P = 5 \times 10^{-34}$) between the 5′- and 3′-UTRs and the presence of G-PQS in the two ends is not independent. This may be suggestive of a functional long-range interaction. Indeed, such long-range interactions between sequences in the 5′- and 3′-UTRs have been proposed to enhance translation efficiency (6,29). We tested whether this correlation was simply due to a correlation in UTR lengths, and found no significant correlation between the length of a 5′-UTR and the corresponding 3′-UTR (Pearson $r = 0.02 \pm 0.04$).

Table 2. Gene families with significantly common G-PQS in either the 5′-or 3′-UTRs

GO code	Description	5′UTR −log p	3′-UTR −log p
GO:0001505	Regulation of neurotransmitter levels	5.9	–
GO:0001996	Positive regulation of heart contraction rate by Epinephrine–norepinephrine	–	5.5
GO:0001997	Increased strength of heart contraction by epinephrine-norepinephrine	–	5.5
GO:0003700	Transcription factor activity	5.8	10.5
GO:0003707	Steroid hormone receptor activity	–	7.4
GO:0004250	Aminopeptidase I activity	6.9	–
GO:0004345	Glucose-6-phosphate 1-dehydrogenase activity	–	5.5
GO:0004385	Guanylate kinase activity	5.9	–
GO:0004409	Homoaconitate hydratase activity	–	5.5
GO:0004965	GABA-B receptor activity	–	6.4
GO:0005083	Small GTPase-regulator activity	5.4	–
GO:0005085	Guanyl-nucleotide exchange factor activity	13.9	12.8
GO:0005089	Rho guanyl-nucleotide exchange factor activity	10.9	9.2
GO:0006003	Fructose 2,6-bisphosphate metabolic process	–	6.6
GO:0006308	DNA catabolic process	–	5.8
GO:0007264	Small GTPase mediated signal transduction	7.1	–
GO:0007275	Multicellular organismal development	–	9.2
GO:0007399	Nervous system development	–	7.0
GO:0016600	Flotillin complex	6.1	–
GO:0035023	Regulation of Rho protein signal transduction	10.3	9.2
GO:0042825	TAP complex	–	5.5
GO:0043565	Sequence-specific DNA binding	6.4	8.5
GO:0045944	Positive regulation of transcription from RNA polymerase II promoter	–	7.1

Where significant over-representation was found, the *P*-value associated with the test is presented as a negative logarithm. Where the test was not significant for one of the UTRs, this is marked as '–'.

Table 3. Gene families with significantly rare G-PQS in either the 5′- or 3′- UTRs

GO code	Description	5′UTR −log p	3′UTR −log p
GO:0000786	Nucleosome	6.5	–
GO:0003735	Structural constituent of ribosome	–	10.2
GO:0004872	Receptor activity	9.0	–
GO:0004984	Olfactory receptor activity	22.3	26.6
GO:0005576	Extracellular region	10.2	–
GO:0005739	Mitochondrion	–	10.0
GO:0005840	Ribosome	–	8.3
GO:0006334	Nucleosome assembly	7.5	6.2
GO:0006412	Translation	6.3	8.2
GO:0006955	Immune response	9.7	7.2
GO:0007186	G-protein coupled receptor protein signaling pathway	18.1	12.0
GO:0007608	Sensory perception of smell	20.0	27.7
GO:0008152	Metabolic process	–	8.5
GO:0042612	MHC class I protein complex	5.9	–
GO:0050896	Response to stimulus	18.3	14.8

Where significant under-representation was found, the *P*-value associated with the test is presented as a negative logarithm. Where the test was not significant for one of the UTRs, this is marked as '–'.

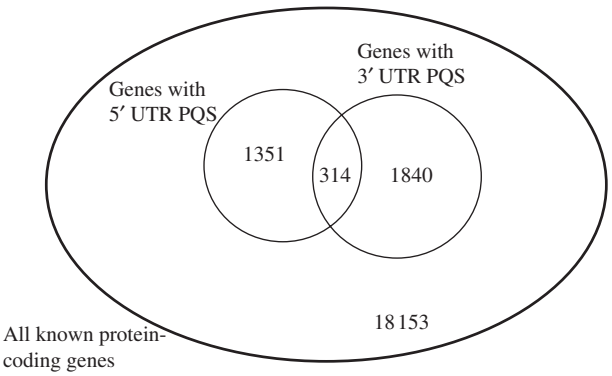

Figure 2. Venn diagram showing the number of genes with G-PQS in their 5′- and/or 3′-UTRs, compared to the number of genes studied. There is a significant overlap between the two sets.

We also confirmed that the same result was found using only non-redundant UTRs (see Table S3).

Positional bias and strand asymmetry of PQS in the 5′-UTR

We then considered whether PQS within the 5′-UTR showed any positional bias. The data presented in Table 4 shows strong clustering of G-PQS in the first ~50 bases, with a gradually declining frequency upon moving further away from the 5′-end. It was noteworthy that there was a strand bias to this positional effect, where C-PQS showed significantly less clustering than G-PQS (significant, $P = 2 \times 10^{-91}$ within 50 bases). We then performed a high-resolution mapping of all PQS in the UTRs, together with the upstream (promoter) region for comparison, studying the frequency with which every base position was occupied by a PQS (Figure 3a). Examination of the 5′-UTR region at this higher resolution (Figure 3a) confirms the results described earlier (Table 4), with clear strand asymmetry and the highest density of G-PQS at the 5′-end of the 5′-UTR, decreasing approximately linearly along the 5′-UTR. In contrast, the C-PQS on the complementary strand are relatively depleted in the first 50 bases at the 5′-end of the 5′-UTR, and then become

Table 4. Frequency of PQS motifs at the 5′-end of the 5′-UTR and the 3′-end of the 3’-UTR

Start	5′-end of 5′-UTR		3′-end of 3’-UTR	
	G-PQS (%)	C-PQS (%)	G-PQS (%)	C-PQS (%)
Within first 10 bases	193 (8.3)	135 (4.4)	13 (0.4)	0 (0)
Within first 20 bases	316 (13.5)	241 (7.9)	18 (0.5)	0 (0)
Within first 50 bases	653 (28.0)	569 (18.5)	106 (3.0)	34 (0.8)
Within first 100 bases	1091 (46.7)	1113 (36.3)	356 (10.1)	266 (5.9)
Within first 1/20th	253 (10.8)	201 (6.5)	102 (2.9)	57 (1.3)
Within first 1/10th	437 (18.7)	356 (11.6)	287 (8.1)	210 (4.7)

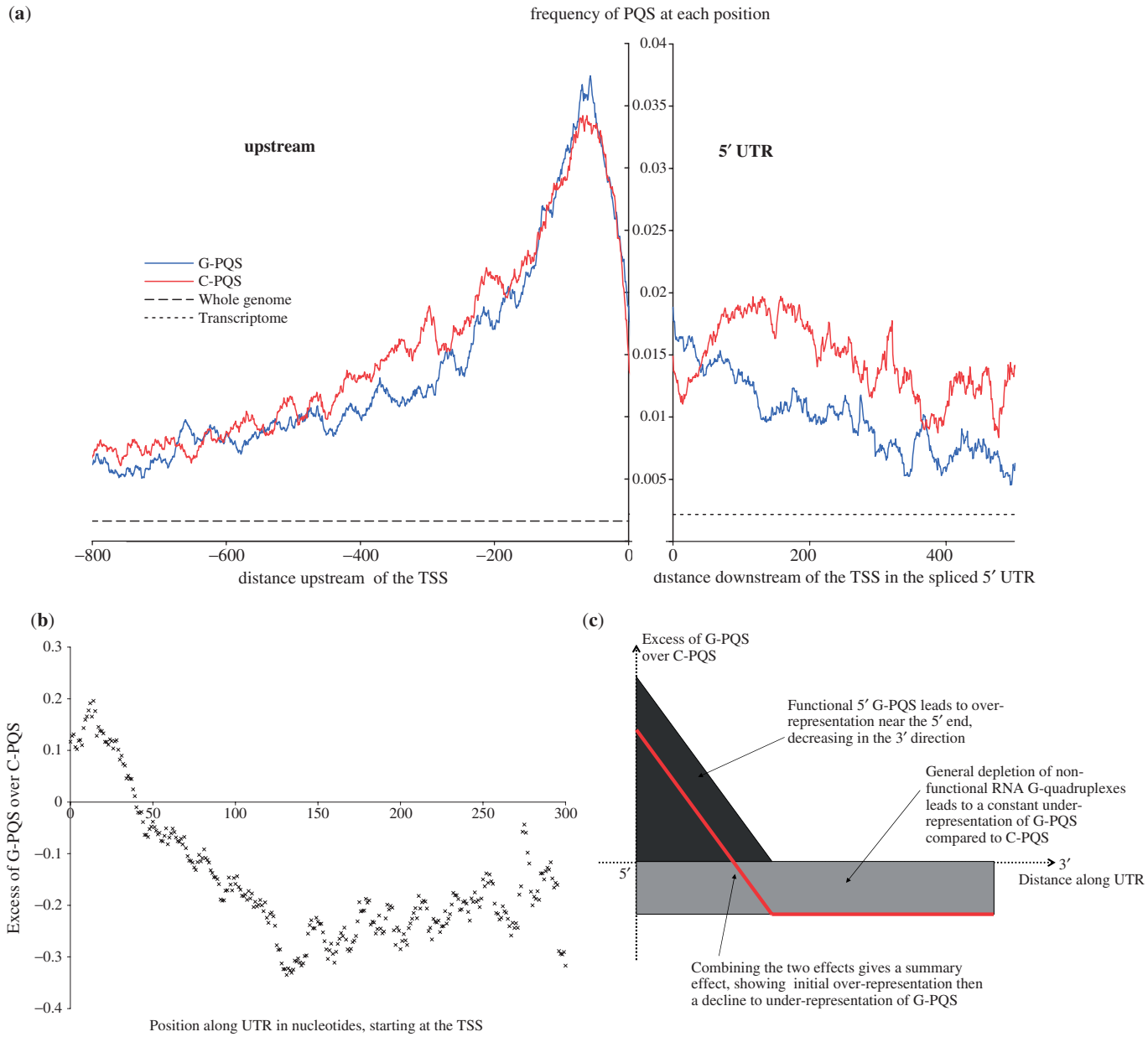

Figure 3. (**a**) High-resolution map of PQS at the TSS junction. Distances are shown in units of bases in the 5′ direction (−ve) or 3′ direction (+ve). The frequencies represent the number of times a PQS was observed to include each individual base position, normalized for the number of times each position was observed. The equivalent frequencies across the whole genome and transcriptome are also shown for comparison. The upstream region is taken from whole genomic data, the 5′-UTR data is taken from the mature mRNA post-splicing, with a gap to highlight the junction. (**b**) Analysis of the excess of G-PQS over C-PQS for 5′-UTRs in the human genome, calculating the excess of G-PQS over C-PQS as Excess = (G-PQS − C-PQS)/(G-PQS + C-PQS) at every position along the 5′-UTR, counting in bases from the TSS. (**c**) A simple model to predict over-/under-representation of G-PQS motifs compared to C-PQS motifs suggests an overall profile given by the red line, in good agreement with the observed data.

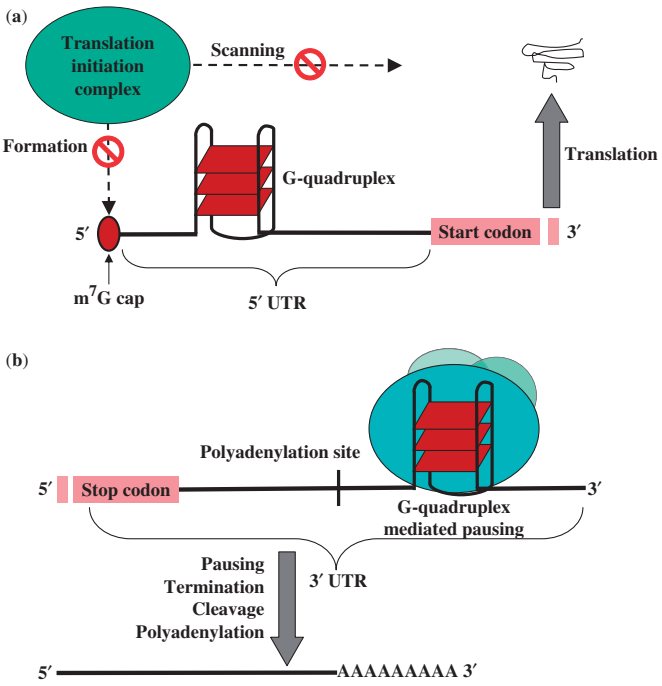

Figure 4. Schematic of hypotheses for the roles of G-PQS associated with UTRs. (**a**) Presence of a stable RNA G-quadruplex, close to the 5′-cap, may prevent translation initiation by sterically blocking the association of the initiation complex at the 5′-cap or by disrupting the scanning process of the small ribosomal subunit towards the start codon. (**b**) G-quadruplex mediated pausing of the transcriptional complex at the 3′-end of the gene may facilitate trancription termination, leading to efficient cleavage at the polyadenylation site and polyadenylation.

increasingly common over the next 50 bases, before decreasing linearly. Over most of the 5′-UTR, and with the striking exception of the first 50 bases at the 5′-end, C-PQS are 30% more common than G-PQS, despite the fact that there are more G than C bases. This strand asymmetry, which is not shared by the PQS in the promoter region, clearly suggests that the functional significance of this sequence may be at the RNA level.

The strand difference can be simply represented at each position along the UTR, by calculating the difference in the proportion of G-PQS and C-PQS divided by the sum of the proportions, yielding a dimensionless and normalized measure of excess. The result of this analysis is shown in Figure 3b. Figure 3c shows a model to describe this strand asymmetry, assuming two factors are affecting the distributions—generalized under-representation of G-PQS resulting from generic deselection in mRNA, and localized over-representation of G-PQS due to a functional role at the 5′-end of the 5′-UTR. This model is in good agreement with the observed asymmetries (Figure 3b). The observed data also contains an interesting periodic variation, with periodicity 30 ± 5 bases; it is unclear as to what this pattern could relate.

It has been previously shown that structural motifs such as hairpins in the 5′-UTRs can modulate mRNA translation efficiency when located in close proximity to the 5′-end of the 5′-UTR, either by interacting directly with the translation machinery (7,30,31) or by acting as target sites for proteins (3,29). We previously showed that a G-quadruplex in the 5′-UTR of the *NRAS* gene significantly reduces translational efficiency (26). Our previous experimental observations (26) coupled with the present computational study lead us to postulate that G-quadruplexes near the 5′-end of 5′-UTRs may be involved in translation regulation (Figure 4a), and hence that there may be an evolutionary pressure in favour of the positional bias of G-quadruplex motifs towards this initial region.

Positional bias and strand asymmetry of PQS in and around the 3′-UTR

We have also carried out a high-resolution mapping of the 3′-UTR and the region immediately downstream of the gene (Figure 5), in analogous fashion to that performed for the 5′-UTRs. The 3′-UTRs of genes have a lower density of PQS than the 5′-UTRs, but there is still a strand bias within the 3′-UTR, with an excess of C-PQS. Two noteworthy features were observed in the immediate vicinity of the transcription end site (TES) junction. First, PQS are extremely rare, compared to all other areas studied, in a region stretching from 20 bases within the 3′-end of the 3′-UTR to 10 bases downstream from the TES. Second, just 3′ of this, there is a very sharp peak in G-PQS density that is not accompanied by an equivalent C-PQS effect (i.e. the strand bias is strong), and is not reflected by an increase in the proportion of G bases (see Supplementary material). These results are suggestive of RNA PQS function localized proximal to the junction between the 3′-UTR and the 3′-downstream region.

In considering the possible functional implications of a G-quadruplex immediately downstream of the TES, we were drawn to observations and suggestions made by Proudfoot and co-workers (32,33) that structures formed by G-rich sequences at the 3′-end of a gene may help to demarcate the end of transcription, especially in cases where there is another gene shortly after the TES. Therefore, given our observation of a peak in G-PQS density in the 100 bases immediately 3′ of the TES, we were prompted to investigate whether known protein-coding genes with G-PQS in this 100 bases region were particularly likely also to have nearby genes. We identified 562 genes with G-PQS in this region, giving a total of 859 such G-PQS. In each case, we measured the distance from these genes to the next known gene in the 3′ direction relative to the gene. In one case, there was no next gene before the end of the chromosome, and that gene was not considered further. The results are shown in Table 5. Of the remaining 561 genes, 50 (8.9%) overlapped with other known genes. Of the remainder, the mean distance to the next gene was 47.3 kb. Of these genes, 97 (17.3%) had another gene within a kilobase. These results were compared to a control set where all known protein-coding genes were considered. Among this dataset of 21 679 genes, 36 did not have a next gene, and 1589 (7.3%) were overlapping. The mean distance to the next gene among the remainder was 93.4 kb, considerably larger than that found for the set with TES-associated G-PQS. Of the entire set of genes, 1633 (7.5%) had another within a kilobase, significantly less than the proportion found for

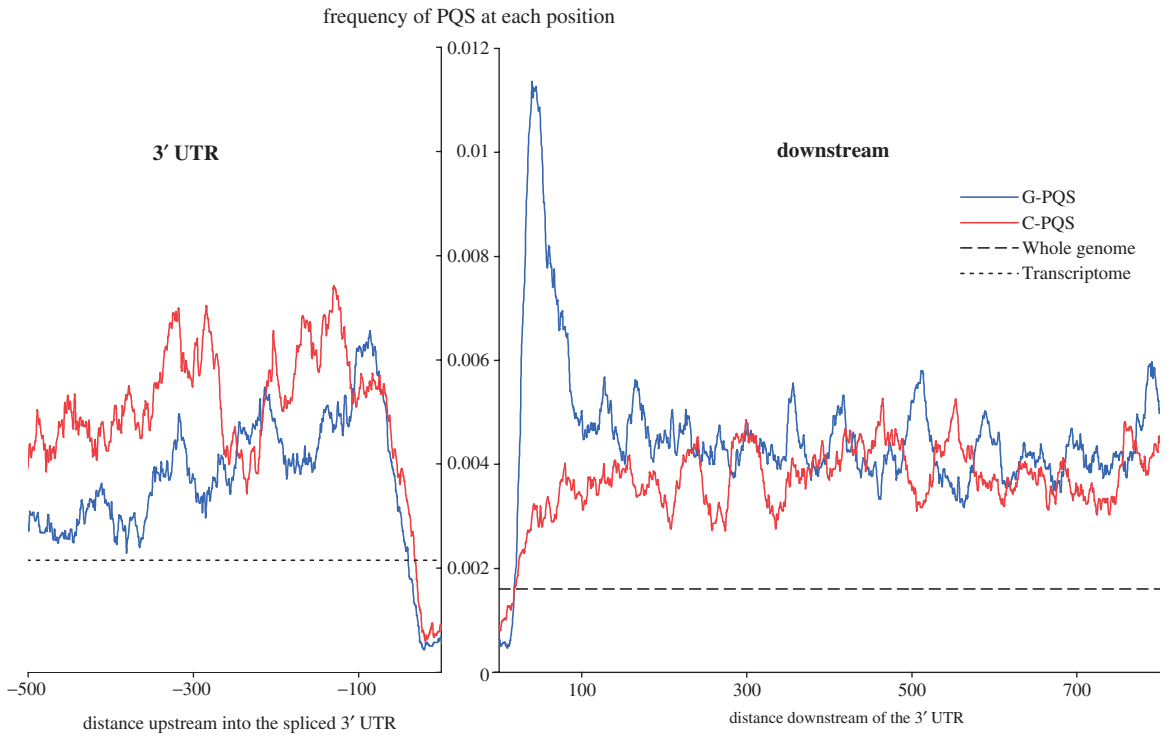

Figure 5. High-resolution map of PQS at the TES junction. Distances are shown in units of bases in the 5′ direction (−ve) or 3′ direction (+ve). The frequencies represent the number of times a PQS was observed to include each individual base position, normalized for the number of times each position was observed. The equivalent frequencies across the whole genome and transcriptome are also shown for comparison. The downstream region is taken from whole genomic data, the 3′-UTR data is taken from the mature mRNA post-splicing, with a gap to highlight the junction.

Table 5. Distance from the 3′-end of genes to the next gene either for all known protein-coding genes with a gene to their 3′-end, or only those with a predicted G-quadruplex immediately to the 3′-end of the gene

All known protein-coding genes	Class	Genes with G-PQS in 100 bases 3′ of TES	Enrichment
No. (%)		No. (%)	
21643 (100)	All with next gene	561 (100)	
1589 (7.3)	Overlapping ($d \leq 0$)	50 (8.9)	1.21
20 054 (92.7)	Non-overlapping ($d > 0$)	511 (91.1)	0.98
1633 (7.5)	Near ($0 < d \leq 1000$)	97 (17.3)	2.29
3376 (15.6)	Medium ($1000 < d \leq 5000$)	114 (20.3)	1.30
15 045 (69.5)	Far ($5000 < d$)	300 (53.5)	0.77
3317 (15.3)	Very far ($100\,000 < d$)	51 (9.1)	0.59

The data is binned into categories where the next gene overlaps the previous one, is within 1 kb ('near', shown in bold), from 1 to 5 kb away ('medium') or >5 kb away ('far'). The subset where the next gene is >100 kb away ('very far') is also shown. Genes where there was no subsequent gene (because they were near a chromosome end) are not shown. Enrichment is calculated as percentage with G-PQS/% of all genes for each class.

genes with TES-associated G-PQS (17.3%), by a factor of 2.3 (significant, $P = 2 \times 10^{-18}$).

A requirement for 3′-end processing of mRNAs is efficient transcription termination (36,37). This appears to be especially important in the cases of closely spaced genes (33). It has been shown that downstream G-rich sequences promote efficient transcription termination. Proudfoot and co-workers (34,35) have previously shown that the sequence $(GGGGGAGGGGG)_4$, a tetramer of four MAZ-binding sites, strongly activates transcription termination *in vitro* when positioned downstream of a synthetic poly(A) site, in a manner that does not require the expression of the MAZ protein. We have noted with interest that this particular G-rich sequence is predicted by *quadparser* to form a stable G-quadruplex (16). Proudfoot and co-workers have further shown that mutation of the sequence to $(GGTGAAAGGTG)_4$, which *quadparser* (16) does not predict to form a G-quadruplex, does not efficiently terminate transcription. Specifically, it was shown *in vivo* that the parent, but not the mutated sequence, promotes efficient transcription termination of a heterologous β-globin construct, and that a naturally occurring G-rich sequence, located 100 nt downstream of the poly (A) site in the human β-actin gene, is essential for transcription termination. Equally, it has been previously noted that some G-rich RNA sequences are involved in

regulating 3′-end processing of alternatively processed mammalian pre-mRNAs by interaction with hnRNP H protein subfamily members (36,37). Our observation that G-PQS cluster immediately after the end of the 3′-UTR supports the proposal that G-quadruplex structure formation could play a role in 3′-end processing of mRNAs by promoting transcription termination and preventing deleterious run-through (Figure 4b).

CONCLUSIONS

From a comprehensive survey of G-quadruplex motifs in and around human genomic UTRs, we have shown that their incidence shows significant strand asymmetry and positional bias, suggestive of functional roles in RNA. In 5′-UTRs, G-quadruplex motifs tend to exist towards the 5′-end of the 5′-UTR supportive of function relating to translation initiation, as depicted in Figure 4a, and for which there is now some experimental support (26). With respect to 3′-UTRs, G-quadruplex motifs tend to cluster immediately after the 3′-end of the mRNA, particularly in cases of genes that have a proximal gene in the 3′ direction. In such cases, failure to terminate transcription could lead to problems associated with the additional transcription of the adjacent gene. We propose that G-quadruplex motifs may serve as pause elements that promote transcriptional termination, cleavage and polyadenylation (Figure 4b).

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

J.L.H. is a Research Councils UK Academic Fellow. We thank the BBSRC for project funding, Cancer Research UK for programme funding and the Cambridge Commonwealth Trust and Trinity College, Cambridge for studentship funding.

FUNDING

Funding for Open Access publication charge: Trinity College, Cambridge.

Conflict of interest statement. None declared.

REFERENCES

1. Pesole,G., Mignone,F., Gissi,C., Grillo,G., Licciulli,F. and Liuni,S. (2001) Structural and functional features of eukaryotic mRNA untranslated regions. *Gene*, **276**, 73–81.
2. Bartel,D.P. (2004) MicroRNAs: genomics, biogenesis, mechanism and function. *Cell*, **116**, 281–297.
3. Wilkie,G.S., Dickson,K.S. and Gray,N.K. (2003) Regulation of mRNA translation by 5′ and 3′-UTR-binding factors. *Trends Bioc. Sci*., **28**, 182–188.
4. Mandal,M. and Breaker,R.R. (2004) Gene regulation by riboswitches. *Nat. Rev. Mol. Cell Biol*., **5**, 451–463.
5. Kozak,M. (1991) Structural features in eukaryotic mRNAs that modulate the initiation of translation. *J. Biol. Chem*., **266**, 19867–19870.
6. Sonenberg,N. (1994) mRNA translation: influence of the 5′ and 3′ untranslated regions. *Curr. Opin. Gen. Dev*., **4**, 310–315.
7. Babendure,J.R., Babendure,J.L., Ding,J.-H. and Tsien,R.Y. (2006) Control of mammalian translation by mRNA structures near caps. *RNA*, **12**, 851–861.
8. Neidle,S. and Balasubramanian,S. (2006) *Quadruplex Nucleic Acids*. RSC, Cambridge.
9. Schaffitzel,C., Berer,I., Postberg,J., Hanes,J., Lipps,H.J. and Plückthun,A. (2001) *In vitro* generated antibodies specific for telomeric guanine-quadruplex DNA react with *Stylonychia lemnae* macronuclei. *Proc. Natl Acad. Sci. USA*, **98**, 8572–8577.
10. Paeschke,K., Simonsson,T., Postberg,J., Rhodes,D. and Lipps,H.J. (2005) Telomere end-binding proteins control the formation of G-quadruplex DNA structures in vivo. *Nat. Struct. Mol. Biol*., **12**, 847–854.
11. Neidle,S. and Parkinson,G.H. (2002) Telomere maintenance as a target for anticancer drug discovery. *Nat. Rev. Drug Disc*., **1**, 383–393.
12. Dexheimer,T.S., Fry,M. and Hurley,L.H. (2006) DNA quadruplexes and gene regulation. In Neidle,S. and Balasubramanian,S. (eds), *Quadruplex Nucleic Acids*. RSC Publishing, Cambridge, UK, pp. 180–207.
13. Siddiqui-Jain,A., Grand,C.L., Bearss,D.J. and Hurley,L.H. (2002) Direct evidence for a G-quadruplex in a promoter region and its targeting with a small molecule to repress *c-MYC* transcription. *Proc. Natl Acad. Sci. USA*, **99**, 11593–11598.
14. Cogoi,S., Quadrifoglio,F. and Xodo,L.E. (2004) G-rich oligonucleotide inhibits the binding of a nuclear protein to the Ki-*ras* promoter and strongly reduces cell growth in human carcinoma pancreatic cells. *Biochemistry*, **43**, 2512–2523.
15. Bejugam,M., Sewitz,S., Shirude,P.S., Rodriguez,R., Shahid,R. and Balasubramanian,S. (2007) Trisubstituted Isoalloxazines as a new class of G-quadruplex binding ligands: small molecule regulation of c-kit oncogene expression. *J. Am. Chem. Soc*., **129**, 12926–12927.
16. Huppert,J.L. and Balasubramanian,S. (2005) Prevalence of quadruplexes in the human genome. *Nucleic Acids Res*., **33**, 2908–2916.
17. Todd,A.K., Johnstone,M. and Neidle,S. (2005) Highly prevalent putative quadruplex sequence motifs in human DNA. *Nucleic Acids Res*., **33**, 2901–2907.
18. Huppert,J.L. and Balasubramanian,S. (2007) G-quadruplexes in promoters throughout the human genome. *Nucleic Acids Res*., **35**, 406–413.
19. Eddy,J. and Maizels,N. (2008) Conserved elements with potential to form polymorphic G-quadruplex structures in the first intron of human genes. *Nucleic Acids Res*., **36**, 1321–1323.
20. Pan,B., Xiong,Y., Shi,K. and Sundaralingam,M. (2003) Crystal structure of a bulged RNA tetraplex at 1.1 Å resolution: implications for a novel binding site in RNA tetraplex. *Structure*, **11**, 1423–1430.
21. Liu,H., Matsugami,A., Katahira,M. and Uesugi,S. (2002) A dimeric RNA quadruplex architecture comprised of two G:G(:A):G:G(:A) hexads, G:G:G:G tetrads and UUUU loops. *J. Mol. Biol*., **322**, 955–970.
22. Christiansen,J., Kofod,M. and Nielsen,F.C. (1994) A guanosine quadruplex and two stable hairpins flank a major cleavage site in insulin-like growth factor II mRNA. *Nucleic Acids Res*., **22**, 5709–5716.
23. Darnell,J.C., Jensen,K.B., Jin,P., Brown,V., Warren,S.T. and Darnell,R.B. (2001) Fragile X mental retardation protein targets G Quartet mRNAs important for neuronal function. *Cell*, **107**, 489–499.
24. Bonnal,S., Schaeffer,C., Creancier,L., Clamens,S., Moine,H., Prats,A.-C. and Vagner,S. (2003) A single internal ribosome entry site containing a G quartet RNA structure drives fibroblast growth factor 2 gene expression at four alternative translation initiation codons. *J. Biol. Chem*., **278**, 39330–39336.
25. Bashkirov,V.I., Scherthan,H., Solinger,J.A., Buerstedde,J.M. and Heyer,W.D. (1997) A mouse cytoplasmic exoribonuclease

(mXRN1p) with preference for G4 tetraplex substrates. *J. Cell Biol.*, **136**, 761–773.
26. Kumari,S., Bugaut,A., Huppert,J.L. and Balasubramanian,S. (2007) An RNA G-quadruplex in the 5′ UTR of the NRAS proto-oncogene modulates translation. *Nat. Chem. Biol.*, **3**, 218–221.
27. Gehring,K., Leroy,J. and Guéron,M. (1993) A tetrameric DNA structure with protonated cytosine·cytosine base pairs. *Nature*, **363**, 499–510.
28. Snoussi,K., Nonin-Lecomte,S. and Leroy,J.-L. (2001) The RNA i-motif. *J. Mol. Biol.*, **309**, 139–153.
29. Preiss,T. and Hentze,M.W. (1999) From factors to mechanisms: translation and translational control in eukaryotes. *Curr. Opin. Gen. Dev.*, **9**, 515–521.
30. Pelletier,J. and Sonenberg,N. (1985) Photochemical cross-linking of cap binding proteins to eucaryotic mRNAs: effect of mRNA 5' secondary structure. *Mol. Cell Biol.*, **5**, 3222–3230.
31. Kozak,M. (1989) Circumstances and mechanisms of inhibition of translation by secondary structure in eucaryotic mRNAs. *Mol. Cell Biol.*, **9**, 5134–5142.
32. Gromak,N., West,S. and Proudfoot,N.J. (2006) Pause sites promote transcriptional termination of mammalian RNA polymerase II. *Mol. Cell Biol.*, **26**, 3986–3996.
33. Ashfield,R., Patel,A.J., Bossone,S.A., Brown,H., Campbell,R.D., Marcu,K.B. and Proudfoot,N.J. (1994) MAZ-dependent termination between closely spaced human complement genes. *EMBO J.*, **13**, 5656–5667.
34. Yonaha,M. and Proudfoot,N.J. (1999) Specific transcriptional pausing activates polyadenylation in a coupled in vitro system. *Mol. Cell*, **3**, 593–600.
35. Yonaha,M. and Proudfoot,N.J. (2000) Transcriptional termination and coupled polyadenylation in vitro. *EMBO J.*, **19**, 3770–3777.
36. Kostadinov,R., Malhotra,N., Viotti,M., Shine,R., D'Antonio,L. and Bagga,P. (2006) GRSDB: a database of quadruplex forming G-rich sequences in alternatively processed mammalian pre-mRNA sequences. *Nucleic Acids Res.*, **34**, D119–D124.
37. Bagga,P., Arhin,G.K. and Wilusz,J. (1998) DSEF-1 is a member of the hnRNP H family of RNA-binding proteins and stimulates pre-mRNA cleavage and polyadenylation in vitro. *Nucleic Acids Res.*, **26**, 5343–5350.