Published online 27 November 2007 *Nucleic Acids Research, 2008, Vol. 36, Database issue* ***D141–D148***
doi:10.1093/nar/gkm982

GRSDB2 and GRS_UTRdb: databases of quadruplex forming G-rich sequences in pre-mRNAs and mRNAs

Oleg Kikin[2], Zachary Zappala[1], Lawrence D'Antonio[2] and Paramjeet S. Bagga[2,]*

[1]Bergen County Academies, Hackensack and [2]Bioinformatics, School of Theoretical and Applied Science, Ramapo College of New Jersey, Mahwah, NJ, USA

Received September 16, 2007; Accepted October 18, 2007

ABSTRACT

G-quadruplex motifs in the RNA play significant roles in key cellular processes and human disease. While sequences capable of forming G-quadruplexes in the pre-mRNA are involved in regulation of polyadenylation and splicing events in mammalian transcripts, the G-quadruplex motifs in the UTRs may help regulate mRNA expression. GRSDB2 is a second-generation database containing information on the composition and distribution of putative Quadruplex-forming G-Rich Sequences (QGRS) mapped in ~29 000 eukaryotic pre-mRNA sequences, many of which are alternatively processed. The data stored in the GRSDB2 is based on computational analysis of NCBI Entrez Gene entries with the help of an improved version of the QGRS Mapper program. The database allows complex queries with a wide variety of parameters, including Gene Ontology terms. The data is displayed in a variety of formats with several additional computational capabilities. We have also developed a new database, GRS_UTRdb, containing information on the composition and distribution patterns of putative QGRS in the 5′- and 3′-UTRs of eukaryotic mRNA sequences. The goal of these experiments has been to build freely accessible resources for exploring the role of G-quadruplex structure in regulation of gene expression at post-transcriptional level. The databases can be accessed at the *G-Quadruplex Resource Site* at: http://bioinformatics.ramapo.edu/GQRS/.

INTRODUCTION

The G-rich polynucleotide molecule can repeatedly fold on itself to form a unimolecular quadruplex structure consisting of stacked G-tetrads, which are square co-planar arrays of four guanine bases each (1). Although G-quadruplexes can also be formed by association of two or four molecules, the present work focuses only on the unimolecular quadruplexes which are more likely to be encountered in physiological conditions (2,3).

G-quadruplexes have come into the limelight in recent years, especially because of increasing indication for their diverse roles in key cellular processes, human disease and as targets for therapy (4–10). Production of G-quadruplexes has been shown to occur cotranscriptionally in the G-rich complementary DNA strands (11). Formation of RNA G-quadruplex structures *in vivo* has also been demonstrated (12). In fact, RNA is more likely to form stable G-quadruplexes than DNA *in vivo* (13,14). G-quadruplex motifs in the RNA have been shown to play significant roles in mRNA turnover (4) and FMRP binding (15). Genes containing FMRP-binding sites may be regulated by a common pathway. At least two such genes have been suggested to be involved in autism (16). We have previously shown that interaction of a G-rich Sequence (GRS) with hnRNPH/H′ can modulate 3′end processing of mammalian pre-mRNAs (17–19). Recently, Furger and coworkers have also found that 3′end processing of melanocortin receptor 1 is regulated by interaction of two G-rich elements with hnRNPH/H′ (20). We have determined, using the QGRS Mapper software program that we had developed earlier (21), that GRS in the above studies are potentially capable of forming stable G-quadruplexes. The hnRNPs H/H′ and F that bind to GRS are known to regulate polyadenylation and splicing events in mammalian transcripts (22–24). The hnRNP A1, which is found in alternative splicing reactions, also has a demonstrated affinity for the G-quadruplex structure (25). G-rich motifs that may fold into quadruplexes in the vicinity of RNA-processing sites act as regulators by interacting with hnRNP A1, H/H′ or F proteins (17–19,24,26,27). The majority of human genes are known to undergo alternative polyadenylation (28), or alternative splicing (29). The role of quadruplex structure in regulating RNA processing, which is an essential component of differential gene expression, needs to be explored.

*To whom correspondence should be addressed. Tel: +1 201 684 7722; Fax: +1 201 684 7637; Email: pbagga@ramapo.edu

The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint Authors.

Prevalence of G-quadruplexes in the human genome has been established (30,31). In a recent study, gene function was found to be associated with potential for G-quadruplex formation (32). However, there is a paucity of systematic studies focusing on the analysis of G-quadruplex motifs near RNA processing sites of the genes, especially that are alternatively processed. Genes that contain G-quadruplex forming sequences are likely to be regulated via special mechanisms (32). Our group has been interested in studying the role of G-quadruplexes in regulation of gene expression at post-transcriptional level. We have adopted a bioinformatics approach to study composition and patterns of G-quadruplexes in pre-mRNA and mRNA sequences.

We had previously built GRSDB, a database of mapped G-quadruplex sequences in selected alternatively processed human and mouse genes (33). GRSDB2 is a second-generation database and contains information on composition and distribution of putative Quadruplex-forming G-Rich Sequences (QGRS) mapped in a large number of eukaryotic pre-mRNA sequences, many of which are alternatively processed (alternatively spliced or alternatively polyadenylated). The data stored in the GRSDB2 is based on computational analysis of NCBI Entrez Gene entries and their corresponding annotated genomic nucleotide sequences of RefSeq/GenBank. GRSDB2 has been built with a new and much improved version of QGRS Mapper program (21). It contains data from ~29 000 eukaryotic genes from other organisms in addition to human and mouse. The data model of GRSDB2 is different than the first version in that it is centered around Entrez Gene rather than solely GenBank/RefSeq nucleotide entries. The search module has been greatly enhanced, making it possible to generate complex queries to search the database with a wide variety of parameters including Gene Ontology terms. The user may select subsets of genes from a query and perform further computations on these genes through a 'Workbench'. It is also possible to define the composition and size of G-quadruplexes to be displayed by applying a variety of filters through the 'options' menu. The 'Gene View', 'Data View' and a highly interactive 'Graphic View' for individual database entries have been significantly enhanced with several additional computational capabilities and links. The data can now be exported into Excel for further analysis. In addition, we have added a 'Sequence View', which displays mapped G-quadruplexes in the context of pre-mRNA sequence.

G-quadruplexes in the mRNA can influence translation initiation (34) as well as repression (35). Recently, a G-quadruplex in the 5′-UTR (untranslated region) of *NRAS* proto-oncogene mRNA was found to inhibit its translation (14). This study also found G-quadruplexes in the 5′-UTR of many other genes. The UTRs of mRNAs contain motifs that are vital for regulation of post-transcriptional gene expression. Much attention has been paid to study the composition of regulatory RNA motifs and mechanism of their interactions with the cellular machinery (36). Our preliminary bioinformatics studies have found notable frequencies of G-quadruplex motifs in the 5′- as well as 3′- UTRs of mammalian mRNAs. More detailed studies are needed to investigate the role of UTR G-quadruplex structure in regulating post-transcriptional gene expression. We have developed a new database, GRS_UTRdb, which contains information on the composition and distribution patterns of putative Quadruplex forming GRS in the 5′- and 3′-UTRs of eukaryotic mRNA sequences. The data stored in the GRS_UTRdb is based on computational analysis of NCBI Entrez Gene entries and their corresponding annotated nucleotide sequences of RefSeq/GenBank. The computations were performed with the help of an extension of the existing QGRS Mapper program (20).

Both the GRSDB2 and GRS_UTRdb databases can be accessed at the G-Quadruplex Resource Site at: http://bioinformatics.ramapo.edu/GQRS/. The goal of these experiments has been to build resources for exploring the role of G-quadruplex structure in regulation of gene expression at post-transcriptional level. Researchers will find both the websites to be user-friendly along with comprehensive help sections as well as context-specific help where it is needed. Investigators interested in the functional relevance of G-quadruplex structure, in particular its role in regulating the gene expression at post-transcriptional level, will find both the databases to be of great value. While GRSDB2 is useful for studying G-quadruplexes near RNA-processing sites, particularly in alternatively processed pre-mRNAs, GRS_UTRdb offers a resource for investigating G-quadruplexes in the untranslated regions of mRNA. Both the websites allow a comprehensive large-scale analysis as well as detailed studies in individual genes.

G-QUADRUPLEX MOTIF

The G-quadruplex motif may be written $G_xN_aG_xN_bG_x N_cG_x$, namely, four guanine groups of equal size (which we call G-groups) interspersed by three arbitrary nucleotide sequences called loops. The size of each G-group corresponds to the number of stacked G-tetrads forming the quadruplex structure. We have previously described the G-quadruplex motif in more detail (21).

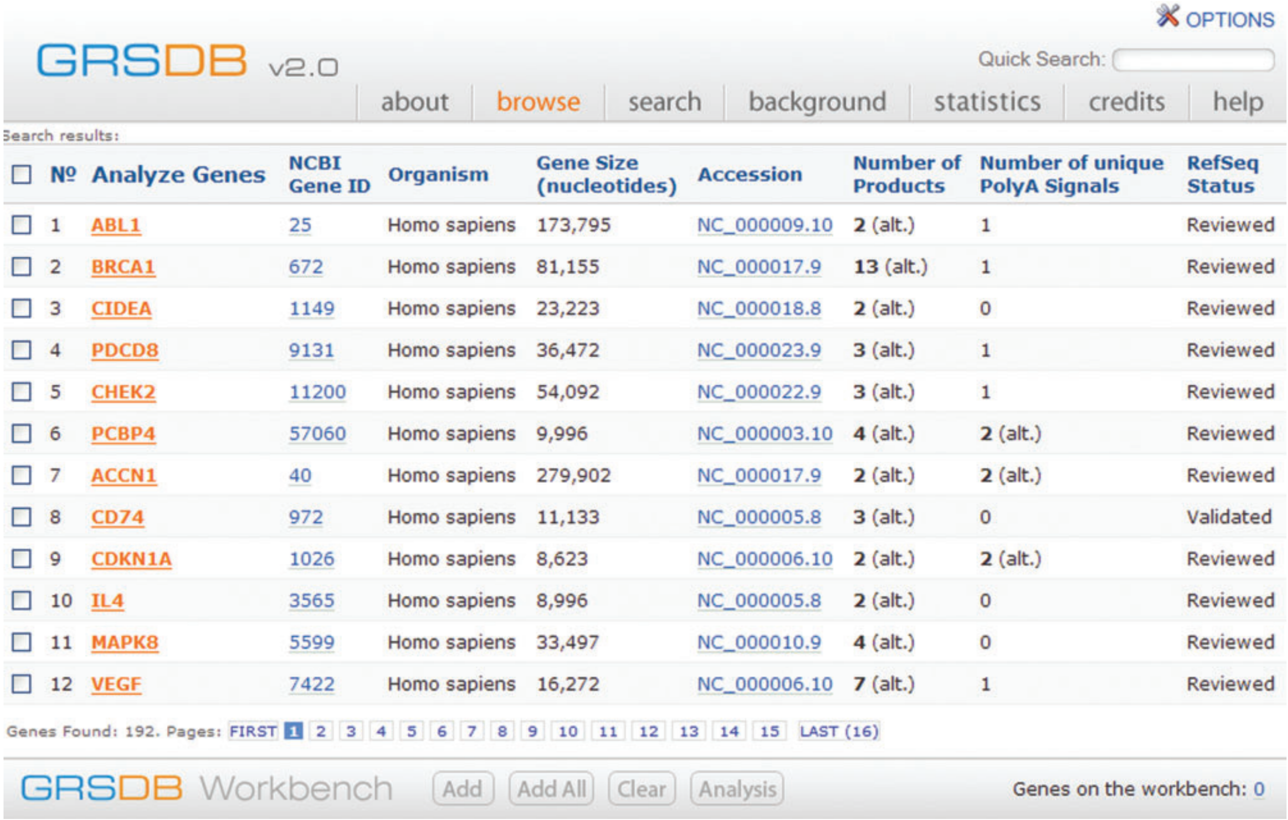
The potential of G-quadruplex to influence gene expression relies on the stability of the structure. Stability of the G-quadruplex is considered to be linked to its loop lengths and the number of G-tetrads in the folded structure (37–39). While quadruplexes with at least three G-tetrads have been accepted as stable structures, two G-tetrad quadruplexes are not uncommon (40,41). In fact, a stable two G-tetrad RNA G-quadruplex that is capable of significantly influencing gene expression *in vivo* has recently been reported (12). Lower stability, in fact may allow more sensitive control of gene expression (12). Two G-tetrads, although relatively lower in stability, are expected to be far more prevalent in the genomes as compared to the three G-tetrads.

METHODS

GRSDB2 and GRS_UTRdb are relational databases developed with MySQL and store non-redundant data.

Table 1. Statistics for GRSDB2

Organism	Number of genes	Alternatively spliced	Averagegene size	Average number of products	Number of (30,2) QGRS	Number of (45,3) QGRS
Homo sapiens	10 475	3197 (30.5%)	61 401	1.54	2 391 014	196 949
Mus musculus	2008	421 (21%)	52 715	1.33	371 809	31 108
Drosophila melanogaster	12 223	3018 (24.7%)	5361	1.46	190 325	7830
Rattus norvegicus	1477	37 (2.5%)	7907	1.04	39 650	2948
Caenorhabditis elegans	3054	1085 (35.5%)	4478	1.54	20 194	309
Gallus gallus	41	0 (0%)	19 878	1	2349	211
Bos taurus	3	0 (0%)	8410	1	201	34
Danio rerio	7	0 (0%)	15 912	1	141	3
Total	29 288	7758			3 015 683	239 392



Search results:

☐	№	Analyze Genes	NCBI Gene ID	Organism	Gene Size (nucleotides)	Accession	Number of Products	Number of unique PolyA Signals	RefSeq Status
☐	1	ABL1	25	Homo sapiens	173,795	NC_000009.10	**2** (alt.)	1	Reviewed
☐	2	BRCA1	672	Homo sapiens	81,155	NC_000017.9	**13** (alt.)	1	Reviewed
☐	3	CIDEA	1149	Homo sapiens	23,223	NC_000018.8	**2** (alt.)	0	Reviewed
☐	4	PDCD8	9131	Homo sapiens	36,472	NC_000023.9	**3** (alt.)	1	Reviewed
☐	5	CHEK2	11200	Homo sapiens	54,092	NC_000022.9	**3** (alt.)	1	Reviewed
☐	6	PCBP4	57060	Homo sapiens	9,996	NC_000003.10	**4** (alt.)	**2** (alt.)	Reviewed
☐	7	ACCN1	40	Homo sapiens	279,902	NC_000017.9	**2** (alt.)	**2** (alt.)	Reviewed
☐	8	CD74	972	Homo sapiens	11,133	NC_000005.8	**3** (alt.)	0	Validated
☐	9	CDKN1A	1026	Homo sapiens	8,623	NC_000006.10	**2** (alt.)	**2** (alt.)	Reviewed
☐	10	IL4	3565	Homo sapiens	8,996	NC_000005.8	**2** (alt.)	0	Reviewed
☐	11	MAPK8	5599	Homo sapiens	33,497	NC_000010.9	**4** (alt.)	0	Reviewed
☐	12	VEGF	7422	Homo sapiens	16,272	NC_000006.10	**7** (alt.)	1	Reviewed

Genes Found: 192. Pages: FIRST 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 LAST (16)

Genes on the workbench: 0

Figure 1. GRSDB2 Query Results Page. Results of a query for alternatively spliced human or rat genes involved in apoptosis. The results may be sorted by clicking the header of any column. At the bottom of the screen, there are four controls allowing the user to add and clear genes from a 'Workbench'. Several programs are provided on the 'Workbench' for further analysis of QGRS from any set of genes in the database.

Interfaces for the databases were built using PHP and Java. The databases have been populated with the help of custom software developed previously by us to analyze NCBI Entrez Gene entries (21). The QGRS are mapped within the relevant gene sequence and assigned a computed value, called a G-score, which rewards those sequences deemed more likely to form a stable complex (21).

Structure and features of GRSDB2

GRSDB2 contains information on the composition and distribution of QGRS mapped in the eukaryotic pre-mRNA sequences.

We have made an effort to include all possible G-quadruplexes. Users may search for QGRS containing G-groups of 2, 3 or more. Also, the length of QGRS and loop size are search parameters that the user may set.

There are two categories of QGRS that are regularly used in GRSDB2: (30,2) refers to QGRS at most 30 nt long and having at least 2 G's per G-group, while (45,3) refers to QGRS at most 45 nt long and having at least 3 G's per G-group.

The overall statistics for the database are shown in Table 1.

Queries may be performed using a variety of search fields. Several fields for gene identifiers are provided: GeneID, Gene Symbol, Gene Name, Aliases, GI and Accession Number. Since alternatively processed genes are a focus of the database, the user may look for genes with a specified number of products and poly(A) signals. The user may also specify which organism(s) to consider.

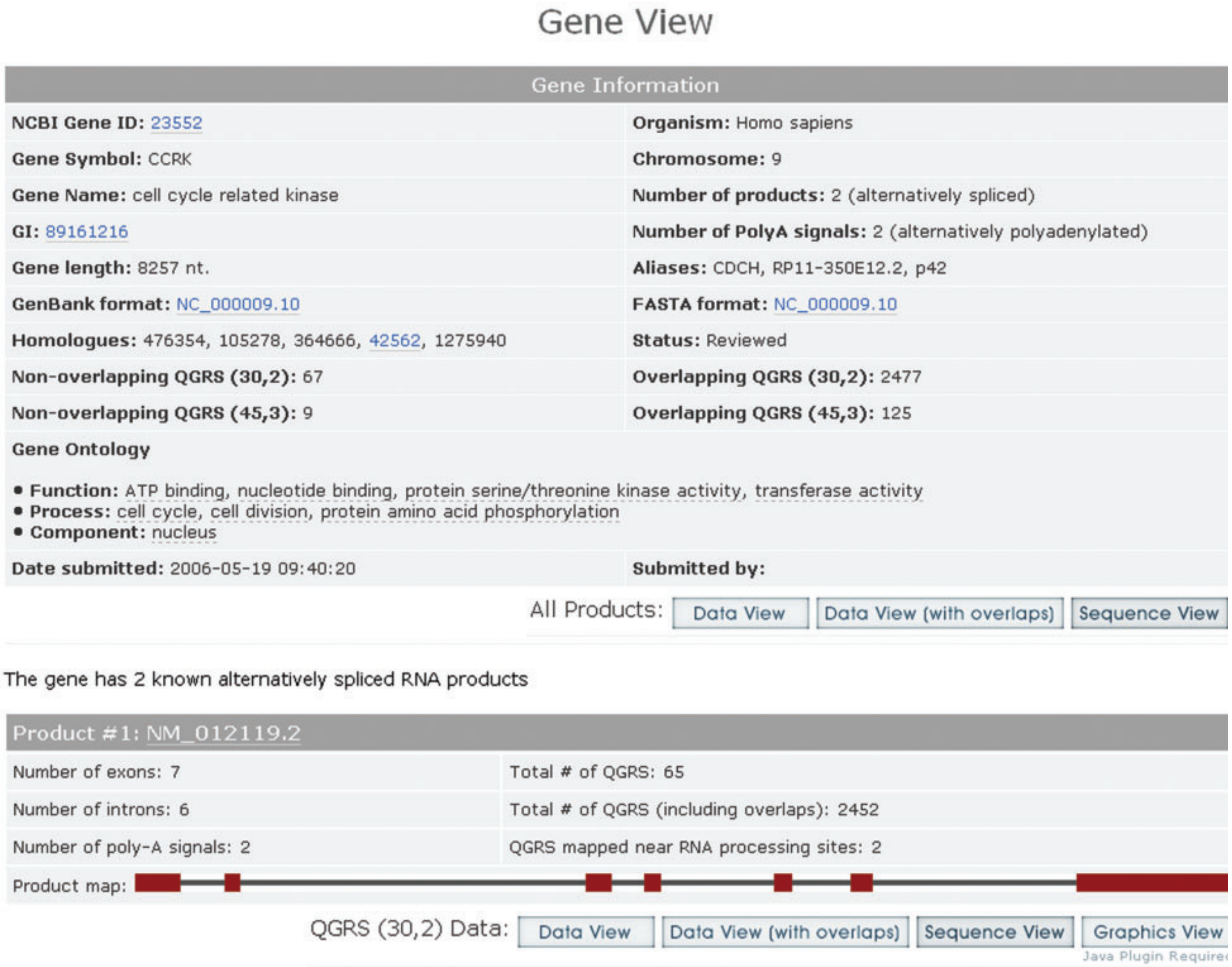
A significant feature of GRSDB2 is the ability it affords the user to look for correlations of occurrences of G-quadruplexes with gene ontology terms. Queries can specify gene ontology function, process or component. The GO terms are an example of a search field for which the user is not required to exactly match the database entry. Instead, searches may be done for which one or more fields start with, end with or contain the query value.

The columns of the query results page consist of the Gene Name, GeneID, Organism, Gene Size, Accession Number, Number of Products and Number of Poly(A) Signals (Figure 1). The results may be sorted on any of these columns, with Gene Name the default sort field. The RefSeq status of each entry is also listed in the table.

The user may analyze sets of genes by putting them into a 'Workbench', where a variety of programs are available to study QGRS distribution patterns for that particular gene set. There are four controls at the bottom of the results page for working with the 'Workbench'. The user may add marked genes, add all genes from the query, clear all genes from the 'Workbench' or analyze the genes on the 'Workbench'.

There are five programs available on the 'Workbench'. One program reports various statistics of the selected genes similar to the statistics page for the entire database. There are two programs [one for (30,2) QGRS, the other for (45,3) QGRS] summarizing the distribution of QGRS with respect to location in exons, introns and near poly(A) signals (which is defined to be within 200 nt). Additionally, there are two programs showing the distribution of G-scores for the selected genes. The user is given the opportunity to export the output of any of these programs to Excel for further analysis.

On the results page, the user may select a particular gene for analysis by clicking on an entry in the Gene Name column. GRSDB2 has five interfaces for viewing information about QGRS: Gene View, Data View (no overlaps), Data View (with overlaps), Sequence View, Graphic View.



Gene View

Gene Information	
NCBI Gene ID: 23552	**Organism:** Homo sapiens
Gene Symbol: CCRK	**Chromosome:** 9
Gene Name: cell cycle related kinase	**Number of products:** 2 (alternatively spliced)
GI: 89161216	**Number of PolyA signals:** 2 (alternatively polyadenylated)
Gene length: 8257 nt.	**Aliases:** CDCH, RP11-350E12.2, p42
GenBank format: NC_000009.10	**FASTA format:** NC_000009.10
Homologues: 476354, 105278, 364666, 42562, 1275940	**Status:** Reviewed
Non-overlapping QGRS (30,2): 67	**Overlapping QGRS (30,2):** 2477
Non-overlapping QGRS (45,3): 9	**Overlapping QGRS (45,3):** 125

Gene Ontology

- **Function:** ATP binding, nucleotide binding, protein serine/threonine kinase activity, transferase activity
- **Process:** cell cycle, cell division, protein amino acid phosphorylation
- **Component:** nucleus

Date submitted: 2006-05-19 09:40:20	**Submitted by:**

The gene has 2 known alternatively spliced RNA products

Product #1: NM_012119.2	
Number of exons: 7	Total # of QGRS: 65
Number of introns: 6	Total # of QGRS (including overlaps): 2452
Number of poly-A signals: 2	QGRS mapped near RNA processing sites: 2

Figure 2. GRSDB2 CCRK Gene View. Provides basic gene information, including the number of products and poly(A) signals, and gene ontology terms for that gene. QGRS counts are displayed for the (30,2) and (45,3) categories together with non-overlapping versus overlapping QGRS. Additional QGRS information together with an exon/intron map is provided for each mRNA product. There are controls to navigate to any of the other views.

The Gene View has a table presenting basic information on the properties of the gene, the number of QGRS found, and the gene ontology terms associated with the gene (Figure 2). For each alternatively spliced product a map of the exon and intron structure is given together with QGRS information for that product.

In the Gene View, the user has several options available to filter which QGRS will be displayed in the Data and Sequence Views. For example, the G-score, loop size, minimum size G-group and maximum QGRS length may be set by the user.

From the Gene View, the user may choose any of the other interfaces. The Data View displays the actual nucleotide sequence for each QGRS and locates its position in an exon, intron or near a poly(A) signal (Figure 3). The results of this page may be exported to Excel. There are two versions of the Data View. One view shows only non-overlapping QGRS, the other view displays all QGRS, overlapping or not.

In the Sequence View, the nucleotide sequence for the entire gene is displayed. Exons are listed in purple and each QGRS is shown in yellow. The Graphic View gives the user a highly interactive visual tool to zoom in on any portion of the gene and analyze QGRS located in that section.

Structure and features of GRS_UTRdb

GRS_UTRdb contains information on the composition and distribution of QGRS in the UTRs of eukaryotic mRNA sequences.

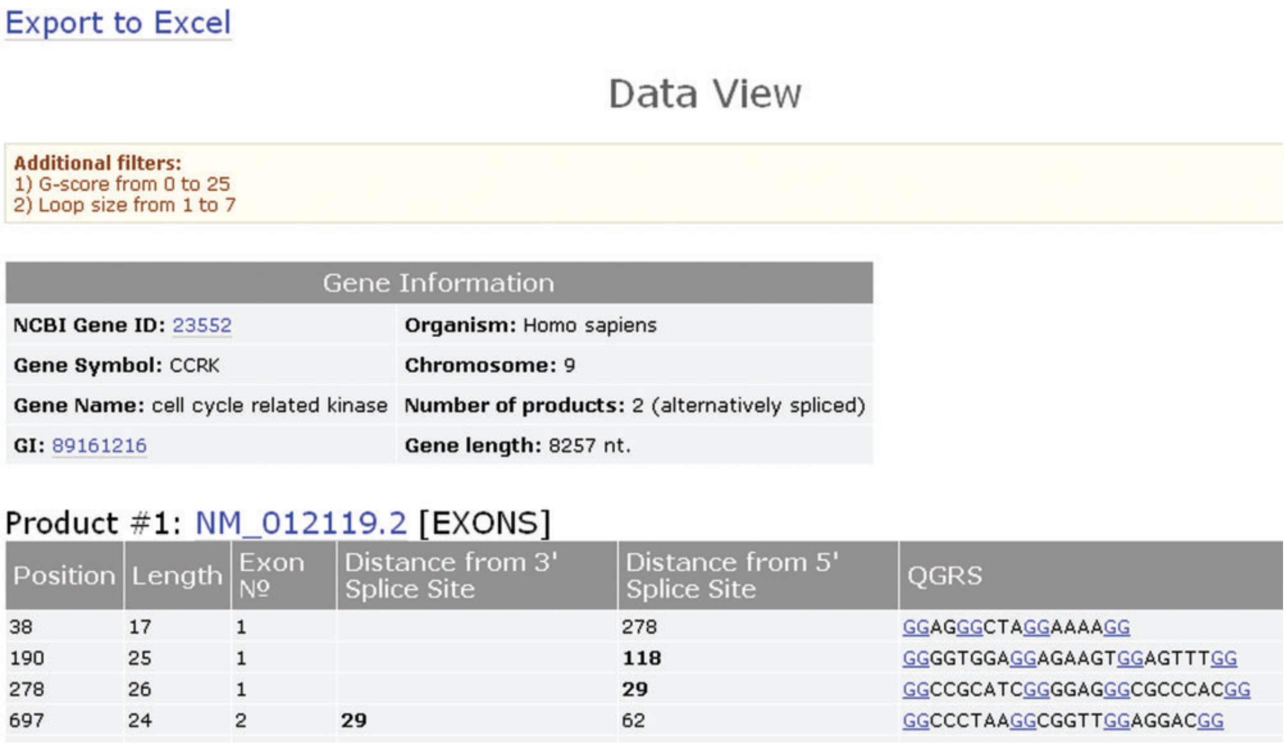
Database users have 14 search fields to define queries. Fields such as GeneID, Gene Name/Symbol, GI and mRNA Accession Number allow the user to look for specific genes and mRNA products. There are fields for gene ontology function, process or component. The user may select which organism to search on. Also, ranges for the lengths of mRNA, 5′ UTR, 3′UTR or CDS may be specified.

Query results are summarized in a table from which the user may select a product for further analysis. GRS_UTRdb has six ways of viewing mRNA data: mRNA Map, Data View (no overlaps), Data View (with overlaps), Sequence View, Gene View, Alternate Products.

The mRNA map contains a table showing an overview of information about the product, including the lengths of the 5′ UTR, CDS and 3′ UTR (Figure 4). The number of QGRS found in each region is displayed. There is a visual map of the locations of QGRS within each region of the product.

The Data View shows the actual sequence of every QGRS in the product. The view gives the location of each QGRS in the 5′ UTR, CDS, 3′ UTR or near poly(A) signal. Also, G-score and the distance of QGRS from the nearest region boundaries are shown. There are two versions of the Data View, one that displays only non-overlapping QGRS and a view that shows all QGRS.

The sequence for the entire product may be found in the Sequence View and separate displays are shown for each region in the product. The QGRS are shown in a box, with each G-group in purple (Figure 5).



Export to Excel

Data View

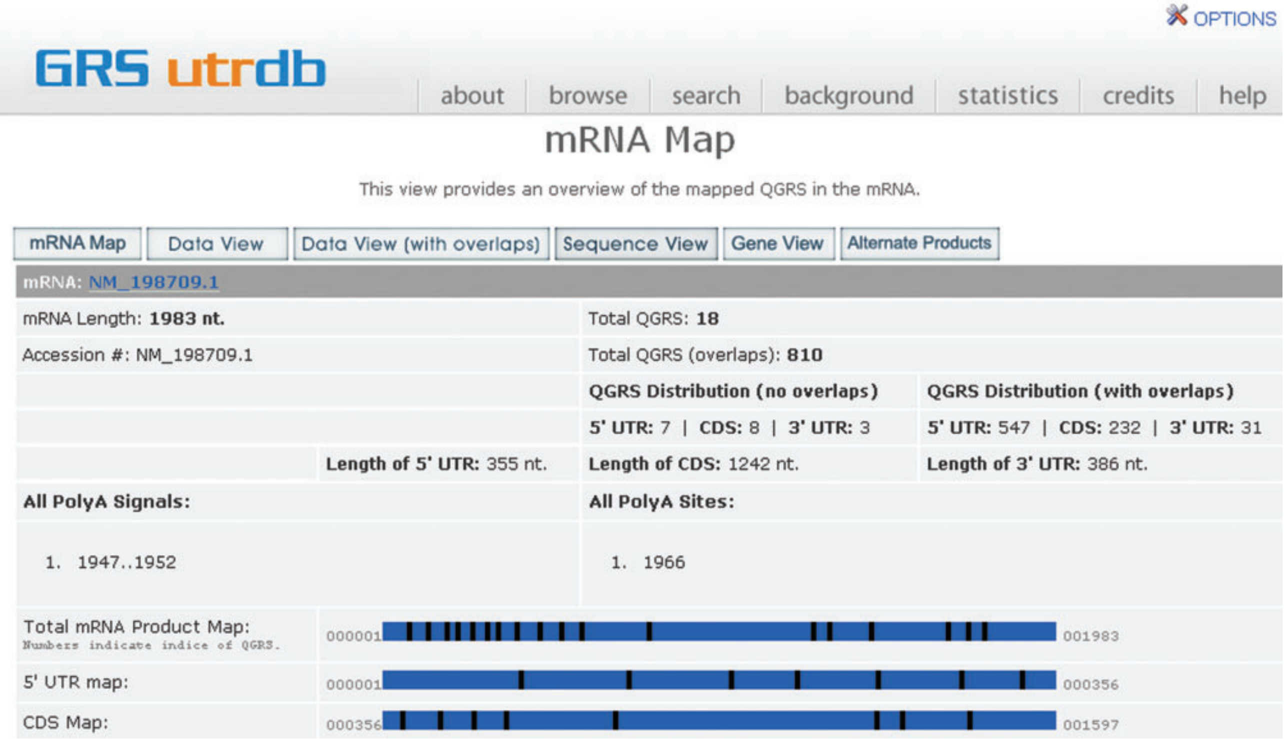
Additional filters:
1) G-score from 0 to 25
2) Loop size from 1 to 7

Gene Information	
NCBI Gene ID: 23552	**Organism:** Homo sapiens
Gene Symbol: CCRK	**Chromosome:** 9
Gene Name: cell cycle related kinase	**Number of products:** 2 (alternatively spliced)
GI: 89161216	**Gene length:** 8257 nt.

Product #1: NM_012119.2 [EXONS]

Position	Length	Exon №	Distance from 3' Splice Site	Distance from 5' Splice Site	QGRS
38	17	1		278	GGAGGGCTAGGAAAAGG
190	25	1		**118**	GGGGTGGAGGAGAAGTGGAGTTTGG
278	26	1		**29**	GGCCGCATCGGGGAGGGCGCCCACGG
697	24	2	**29**	62	GGCCCTAAGGCGGTTGGAGGACGG

Figure 3. GRSDB2 CCRK Data View. A listing of the nucleotide sequences for all QGRS, the location of QGRS in exons, introns and near poly(A) signals. Results are shown for each product of the gene. What is shown is a truncated image of the view. The QGRS displayed satisfy the conditions that G-scores are in the range from 0 to 25 (which in effect restricts the output to QGRS with G-groups of size 2) and loop size is in the range from 1 to 7.



mRNA Map

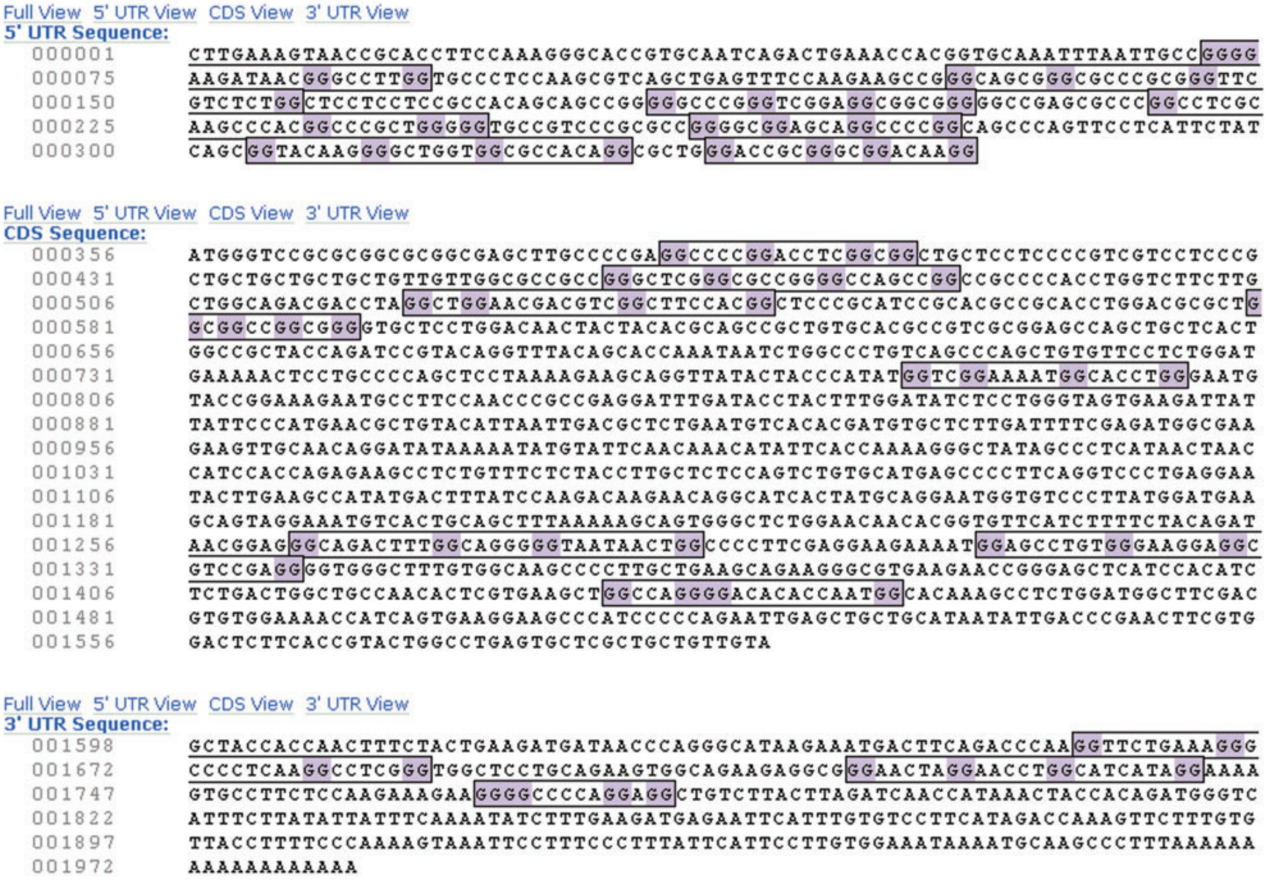
This view provides an overview of the mapped QGRS in the mRNA.

mRNA: NM_198709.1

mRNA Length: **1983 nt.**		Total QGRS: **18**	
Accession #: NM_198709.1		Total QGRS (overlaps): **810**	
		QGRS Distribution (no overlaps)	**QGRS Distribution (with overlaps)**
		5' UTR: 7 | **CDS:** 8 | **3' UTR:** 3	**5' UTR:** 547 | **CDS:** 232 | **3' UTR:** 31
	Length of 5' UTR: 355 nt.	**Length of CDS:** 1242 nt.	**Length of 3' UTR:** 386 nt.
All PolyA Signals:		**All PolyA Sites:**	
1. 1947..1952		1. 1966	

Total mRNA Product Map:
Numbers indicate indice of QGRS.

Figure 4. GRS_UTRdb NM_198709.1 mRNA (ARSB Gene) Map. Displays basic information about the 5′ UTR, CDS and 3′ UTR and the QGRS frequency in each region. A visual map of each region is displayed with the relative positions of QGRS depicted in the map.



5' UTR Sequence:

```
000001    CTTGAAAGTAACCGCACCTTCCAAAGGGCACCGTGCAATCAGACTGAAACCACGGTGCAAATTTAATTGCCGGGG
000075    AAGATAACGGGCCTTGGTGCCCTCCAAGCGTCAGCTGAGTTTCCAAGAAGCCGGGCAGCGGGCGCCCGCGGGTTC
000150    GTCTCTGGCTCCTCCTCCGCCACAGCAGCCGGGGGCCCGGGTCGGAGGCGGCGGGGGCCGAGCGCCCGGCCTCGC
000225    AAGCCCACGGCCCGCTGGGGGTGCCGTCCCGCGCCGGGGCGGAGCAGGCCCCGGCAGCCCAGTTCCTCATTCTAT
000300    CAGCGGTACAAGGGGCTGGTGGCGCCACAGGCGCTGGGACCGCGGGCGGACAAGG
```

CDS Sequence:

```
000356    ATGGGTCCGCGCGGCGCGGCGAGCTTGCCCCGAGGCCCCGGACCTCGGCGGCTGCTCCTCCCCGTCGTCCTCCCG
000431    CTGCTGCTGCTGCTGTTGTTGGCGCCGCCGGGCTCGGGCGCCGGGGCCAGCCGGCCGCCCCACCTGGTCTTCTTG
000506    CTGGCAGACGACCTAGGCTGGAACGACGTCGGCTTCCACGGCTCCCGCATCCGCACGCCGCACCTGGACGCGCTG
000581    GCGGCCGGCGGGGTGCTCCTGGACAACTACTACACGCAGCCGCTGTGCACGCCGTCGCGGAGCCAGCTGCTCACT
000656    GGCCGCTACCAGATCCGTACAGGTTTACAGCACCAAATAATCTGGCCCTGTCAGCCCAGCTGTGTTCCTCTGGAT
000731    GAAAAACTCCTGCCCCAGCTCCTAAAAGAAGCAGGTTATACTACCCATATGGTCGGAAAATGGCACCTGGGAATG
000806    TACCGGAAAGAATGCCTTCCAACCCGCCGAGGATTTGATACCTACTTTGGATATCTCCTGGGTAGTGAAGATTAT
000881    TATTCCCATGAACGCTGTACATTAATTGACGCTCTGAATGTCACACGATGTGCTCTTGATTTTCGAGATGGCGAA
000956    GAAGTTGCAACAGGATATAAAAATATGTATTCAACAAACATATTCACCAAAAGGGCTATAGCCCTCATAACTAAC
001031    CATCCACCAGAGAAGCCTCTGTTTCTCTACCTTGCTCTCCAGTCTGTGCATGAGCCCCTTCAGGTCCCTGAGGAA
001106    TACTTGAAGCCATATGACTTTATCCAAGACAAGAACAGGCATCACTATGCAGGAATGGTGTCCCTTATGGATGAA
001181    GCAGTAGGAAATGTCACTGCAGCTTTAAAAAGCAGTGGGCTCTGGAACAACACGGTGTTCATCTTTTCTACAGAT
001256    AACGGAGGGCAGACTTTGGCAGGGGGTAATAACTGGCCCCTTCGAGGAAGAAAATGGAGCCTGTGGGAAGGAGGC
001331    GTCCGAGGGGTGGGCTTTGTGGCAAGCCCCTTGCTGAAGCAGAAGGGCGTGAAGAACCGGGAGCTCATCCACATC
001406    TCTGACTGGCTGCCAACACTCGTGAAGCTGGCCAGGGGACACACCAATGGCACAAAGCCTCTGGATGGCTTCGAC
001481    GTGTGGAAAACCATCAGTGAAGGAAGCCCATCCCCCAGAATTGAGCTGCTGCATAATATTGACCCGAACTTCGTG
001556    GACTCTTCACCGTACTGGCCTGAGTGCTCGCTGCTGTTGTA
```

3' UTR Sequence:

```
001598    GCTACCACCAACTTTCTACTGAAGATGATAACCCAGGGCATAAGAAATGACTTCAGACCCAAGGTTCTGAAAGGG
001672    CCCCTCAAGGCCTCGGGTGGCTCCTGCAGAAGTGGCAGAAGAGGCGGGAACTAGGAACCTGGCATCATAGGAAAA
001747    GTGCCTTCTCCAAGAAAGAAGGGGCCCCAGGAGGCTGTCTTACTTAGATCAACCATAAACTACCACAGATGGGTC
001822    ATTTCTTATATTATTTCAAAATATCTTTGAAGATGAGAATTCATTTGTGTCCTTCATAGACCAAAGTTCTTTGTG
001897    TTACCTTTTCCCAAAAGTAAATTCCTTTCCCTTTATTCATTCCTTGTGGAAATAAAATGCAAGCCCTTTAAAAAA
001972    AAAAAAAAAAAA
```

Figure 5. GRS_UTRdb NM_198709.1 Sequence View. Shows the nucleotide sequence for the entire product, in separate displays for the 5′ UTR, CDS and 3′ UTR. QGRS are enclosed in a box with each G-group shown in purple.

The Gene View gives basic information about genes, including gene ontology terms and the location of poly(A) signals and sites. Each mRNA product is listed together with the distribution of QGRS there. The 'Alternate Products' tab takes the user directly to the QGRS information for each product associated with the gene.

CONCLUSIONS

GRSDB2 and GRS_UTRdb provide curated data on the composition and distribution of putative QGRS in the transcribed regions of a large number of alternatively processed eukaryotic genes. GRSDB2 is useful for studying G-quadruplexes near RNA-processing sites particularly those that are differentially processed. At present, it contains data for 29 288 genes encompassing 42 932 products from several eukaryotic organisms. More than 3 million QGRS have been mapped to these genes. The availability of large number of pre-mRNAs with mapped QGRS makes it possible to perform a variety of bioinformatics studies. The database website already offers a range of computational tools to aid large scale as well as individual gene analysis. The 'Workbench' can be used to perform computations on sets and sub-sets of genes in the database. The 'Gene View', 'Data View' and 'Sequence View' are useful for studying individual genes and their multiple products. The highly interactive 'Graphic View' is particularly useful for working with parts of the genes.

The new GRS_UTRdb offers a valuable resource for investigating G-quadruplexes in the UTRs of mRNA. Currently, it contains data for more than 16,000 eukaryotic mRNAs, including ~27,000 QGRS which have been mapped to the 5′ UTRs. Like GRSDB2, GRS_UTRdb also displays QGRS data in a variety of modes with computational capabilities and links. At this point, it does not have a 'Workbench' facility. We are constantly adding new genes and new computational tools to the website. Since genes containing G-quadruplex motifs could be regulated through special mechanisms, one can expect for the gene function to correlate with G-quadruplex formation (32). We have classified the gene entries in our databases according to the gene ontology categories, which allows for queries with relevant terms.

Researchers interested in the functional relevance of G-quadruplex structure, in particular its role in regulating the gene expression at post-transcriptional level, will find both the databases to be of great value.

ACKNOWLEDGEMENTS

We thank Manuel Viotti for assistance with the testing and uploading data in the initial stages of this project. The authors would like to acknowledge Rumen Kostadinov, who developed the first version of GRSDB. We would also like to acknowledge Marcelo Halpern for technical assistance with the web server. This project was funded in part by a grant from the Provost Office of Ramapo College of New Jersey. Funding for the Open Access publication charges for this article was provided in part by the Divisions of Student Affairs and Academic Affairs of Ramapo College of New Jersey, and the Bergen County Academies, Hackensack, New Jersey.

Conflict of interest statement. None declared.

REFERENCES

1. Gellert,M., Lipsett,M.N. and Davies,D.R. (1962) Helix formation by guanylic acid. *Proc. Natl Acad. Sci. USA*, **48**, 2013–2018.
2. Schaffitzel,C., Berger,I., Postberg,J., Hanes,J., Lipps,H.J. and Pluckthun,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.
3. Halder,K. and Chowdhury,S. (2005) Kinetic resolution of bimolecular hybridization versus intramolecular folding in nucleic acids by surface plasmon resonance: application to G-quadruplex/duplex competition in human c-myc promoter. *Nucleic Acids Res.*, **33**, 4466–4474.
4. Simonsson,T. (2001) G-quadruplex DNA structures – variations on a theme. *Biol. Chem.*, **382**, 621–628.
5. Davis,J.T. (2004) G-quartets 40 years later: from 5′-GMP to molecular biology and supramolecular chemistry. *Angew. Chem. Int. Ed. Engl.*, **43**, 668–698.
6. Kelland,L.R. (2005) Overcoming the immortality of tumour cells by telomere and telomerase based cancer therapeutics – current status and future prospects. *Eur. J. Cancer*, **41**, 971–979.
7. Burge,S., Parkinson,G.N., Hazel,P., Todd,A.K. and Neidle,S. (2006) Quadruplex DNA: sequence, topology and structure. *Nucleic Acids Res.*, **34**, 5402–5415.
8. Maizels,N. (2006) Dynamic roles for G4 DNA in the biology of eukaryotic cells. *Nat. Struct. Mol. Biol.*, **13**, 1055–1059.
9. 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.
10. Todd,A.K., Haider,S.M., Parkinson,G.N. and Neidle,S. (2007) Sequence occurrence and structural uniqueness of a G-quadruplex in the human c-kit promoter. *Nucleic Acids Res*, **35**, 5799–5808.
11. Duquette,M.L., Handa,P., Vincent,J.A., Taylor,A.F. and Maizels,N. (2004) Intracellular transcription of G-rich DNAs induces formation of G-loops, novel structures containing G4 DNA. *Genes Dev.*, **18**, 1618–1629.
12. Wieland,M. and Hartig,J.S. (2007) RNA quadruplex-based modulation of gene expression. *Chem. Biol.*, **14**, 757–763.
13. Sacca,B., Lacroix,L. and Mergny,J.L. (2005) The effect of chemical modifications on the thermal stability of different G-quadruplex-forming oligonucleotides. *Nucleic Acids Res.*, **33**, 1182–1192.
14. 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.
15. 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.
16. Nishimura,Y., Martin,C.L., Vazquez-Lopez,A., Spence,S.J., Alvarez-Retuerto,A.I., Sigman,M., Steindler,C., Pellegrini,S., Schanen,N.C. *et al.* (2007) Genome-wide expression profiling of lymphoblastoid cell lines distinguishes different forms of autism and reveals shared pathways. *Hum. Mol. Genet.*, **16**, 1682–1698.
17. Bagga,P.S., Ford,L.P., Chen,F. and Wilusz,J. (1995) The G-rich auxiliary downstream element has distinct sequence and position requirements and mediates efficient 3′ end pre-mRNA processing through a trans-acting factor. *Nucleic Acids Res.*, **23**, 1625–1631.
18. Bagga,P.S., 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.
19. Arhin,G.K., Boots,M., Bagga,P.S., Milcarek,C. and Wilusz,J. (2002) Downstream sequence elements with different affinities for the hnRNP H/H′ protein influence the processing efficiency of mammalian polyadenylation signals. *Nucleic Acids Res.*, **30**, 1842–1850.

20. Dalziel,M., Nunes,N.M. and Furger,A. (2007) Two G-rich regulatory elements located adjacent to and 440 nucleotides downstream of the core poly(A) site of the intronless melanocortin receptor 1 gene are critical for efficient 3′ end processing. *Mol Cell Biol.*, **27**, 1568–1580.
21. Kikin,O., D'Antonio,L. and Bagga,P.S. (2006) QGRS Mapper: a web-based server for predicting G-quadruplexes in nucleotide sequences. *Nucleic Acids Res.*, **34**, W676–W682.
22. Veraldi,K.L., Arhin,G.K., Martincic,K., Chung-Ganster,L.H., Wilusz,J. and Milcarek,C. (2001) hnRNP F influences binding of a 64-kilodalton subunit of cleavage stimulation factor to mRNA precursors in mouse B cells. *Mol. Cell. Biol.*, **21**, 1228–1238.
23. Bruce,S.R., Dingle,R.W. and Peterson,M.L. (2003) B-cell and plasma-cell splicing differences: a potential role in regulated immunoglobulin RNA processing. *RNA*, **9**, 1264–1273.
24. Garneau,D., Revil,T., Fisette,J.F. and Chabot,B. (2005) Heterogeneous nuclear ribonucleoprotein F/H proteins modulate the alternative splicing of the apoptotic mediator Bcl-x. *J. Biol. Chem.*, **280**, 22641–22650.
25. Zhang,Q.S., Manche,L., Xu,R.M. and Krainer,A.R. (2006) hnRNP A1 associates with telomere ends and stimulates telomerase activity. *RNA*, **12**, 1116–1128.
26. Han,K., Yeo,G., An,P., Burge,C.B. and Grabowski,P.J. (2005) A combinatorial code for splicing silencing: UAGG and GGGG motifs. *PLoS Biol.*, **3**, e158.
27. Wang,E., Dimova,N. and Cambi,F. (2007) PLP/DM20 ratio is regulated by hnRNPH and F and a novel G-rich enhancer in oligodendrocytes. *Nucleic Acids Res.*, **35**, 4164–4178.
28. Tian,B., Hu,J., Zhang,H. and Lutz,C.S. (2005) A large-scale analysis of mRNA polyadenylation of human and mouse genes. *Nucleic Acids Res.*, **33**, 201–212.
29. Johnson,J.M., Castle,J., Garrett-Engele,P., Kan,Z., Loerch,P.M., Armour,C.D., Santos,R., Schadt,E.E., Stoughton,R. *et al.* (2003) Genome-wide survey of human alternative pre-mRNA splicing with exon junction microarrays. *Science*, **302**, 2141–2144.
30. Huppert,J.L. and Balasubramanian,S. (2005) Prevalence of quadruplexes in the human genome. *Nucleic Acids Res.*, **33**, 2908–2916.
31. Todd,A.K., Johnston,M. and Neidle,S. (2005) Highly prevalent putative quadruplex sequence motifs in human DNA. *Nucleic Acids Res.*, **33**, 2901–2907.
32. Eddy,J. and Maizels,N. (2006) Gene function correlates with potential for G4 DNA formation in the human genome. *Nucleic Acids Res.*, **34**, 3887–3896.
33. 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.
34. 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.
35. Oliver,A.W., Bogdarina,I., Schroeder,E., Taylor,I.A. and Kneale,G.G. (2000) Preferential binding of fd gene 5 protein to tetraplex nucleic acid structures. *J. Mol. Biol.*, **301**, 575–584.
36. Mignone,F., Gissi,C., Liuni,S. and Pesole,G. (2002) Untranslated regions of mRNAs. *Genome Biol.*, **3**, reviews0004.1-reviews0004.10.
37. Crnugelj,M., Sket,P. and Plavec,J. (2003) Small change in a G-rich sequence, a dramatic change in topology: new dimeric G-quadruplex folding motif with unique loop orientations. *J. Am. Chem. Soc.*, **125**, 7866–7871.
38. Hazel,P., Huppert,J., Balasubramanian,S. and Neidle,S. (2004) Loop-length-dependent folding of G-quadruplexes. *J. Am. Chem. Soc.*, **126**, 16405–16415.
39. Risitano,A. and Fox,K.R. (2004) Influence of loop size on the stability of intramolecular DNA quadruplexes. *Nucleic Acids Res.*, **32**, 2598–2606.
40. Zarudnaya,M.I., Kolomiets,I.M., Potyahaylo,A.L. and Hovorun,D.M. (2003) Downstream elements of mammalian pre-mRNA polyadenylation signals: primary, secondary and higher-order structures. *Nucleic Acids Res.*, **31**, 1375–1386.
41. Kankia,B.I., Barany,G. and Musier-Forsyth,K. (2005) Unfolding of DNA quadruplexes induced by HIV-1 nucleocapsid protein. *Nucleic Acids Res.*, **33**, 4395–4403.