Published online 21 September 2010 *Nucleic Acids Research, 2010, Vol. 38, No. 22* ***8149–8163***
doi:10.1093/nar/gkq804

RNA G-Quadruplexes in the model plant species *Arabidopsis thaliana*: prevalence and possible functional roles

Melissa A. Mullen[1], Kalee J. Olson[1,2], Paul Dallaire[3], François Major[3], Sarah M. Assmann[2,*] and Philip C. Bevilacqua[1,*]

[1]Department of Chemistry, [2]Department of Biology, The Pennsylvania State University, University Park, Pennsylvania 16802-5302, USA and [3]Institute for Research in Immunology and Cancer (IRIC), Department of Computer Science and Operations Research, Université de Montréal, PO Box 6128, Downtown Station, Montréal, Québec H3C 3J7, Canada

Received July 15, 2010; Revised August 24, 2010; Accepted August 30, 2010

ABSTRACT

Tandem stretches of guanines can associate in hydrogen-bonded arrays to form G-quadruplexes, which are stabilized by K^+ ions. Using computational methods, we searched for G-Quadruplex Sequence (GQS) patterns in the model plant species *Arabidopsis thaliana*. We found ~1200 GQS with a G_3 repeat sequence motif, most of which are located in the intergenic region. Using a Markov modeled genome, we determined that GQS are significantly underrepresented in the genome. Additionally, we found ~43 000 GQS with a G_2 repeat sequence motif; notably, 80% of these were located in genic regions, suggesting that these sequences may fold at the RNA level. Gene Ontology functional analysis revealed that GQS are overrepresented in genes encoding proteins of certain functional categories, including enzyme activity. Conversely, GQS are underrepresented in other categories of genes, notably those for non-coding RNAs such as tRNAs and rRNAs. We also find that genes that are differentially regulated by drought are significantly more likely to contain a GQS. CD-detected K^+ titrations performed on representative RNAs verified formation of quadruplexes at physiological K^+ concentrations. Overall, this study indicates that GQS are present at unique locations in *Arabidopsis* and that folding of RNA GQS may play important roles in regulating gene expression.

INTRODUCTION

Both DNA and RNA can form G-quartets and G-quadruplexes (1,2), which consist of four guanine bases associated in a planar orientation and stabilized by Hoogsteen-to-Watson–Crick interactions and centrally positioned K^+ or Na^+ ions (Figure 1a) (3–5). Several conformations can be adopted by G-quadruplexes, including parallel and antiparallel topologies, which can be formed in either intramolecular or intermolecular strands. Parallel G-quadruplexes are characterized by parallel orientation of all four strands (Figure 1b) and have anti conformation of all four guanosines, while antiparallel G-quadruplexes are characterized by alternating orientation of strands (Figure 1c) and have two *anti* and two *syn* guanosines per quartet. Conformation of the quadruplex is controlled by strand orientation, loop length and ions present in solution (6,7). Parallel quadruplex structures are preferred by RNA (8), which has been explained by RNA avoiding topologies that would require the *syn* conformation of the nucleosides and their associated C2′-endo sugar puckers (9–11).

Raising the concentration of monovalent cations lowers the free energy of quadruplex formation. While G-quartet forming sequence (GQS) form in the presence of both K^+ and Na^+ ions, they are typically more stable in the presence of K^+ owing to its larger ionic radius, which is associated with a smaller free energy of dehydration (1,12,13). In addition, molecular crowding by osmolytes, including both ions and organic solutes, favors quadruplex stability, owing to the quadruplex's compact and less hydrated structure, and simultaneously disfavors competing Watson–Crick structures (14–18). Given these behaviors, formation of GQS might be favored during

*To whom correspondence should be addressed. Tel: +1 814 863 3812; Fax: +1 814 865 2927; Email: pcb@chem.psu.edu
Correspondence may also be addressed to Sarah M. Assmann. Tel: +1 814 863 9579; Fax: +1 814 865 9131; Email: sma3@psu.edu

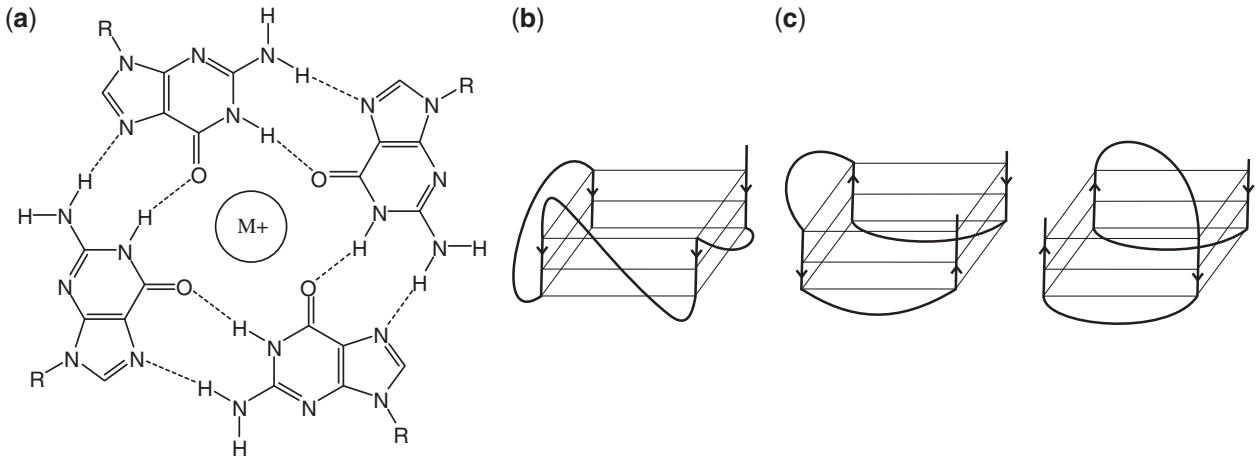

Figure 1. G-quartet and G-quadruplex structures and topologies. (**a**) G-quartet structure, showing Hoogsteen-to-Watson-Crick face hydrogen bonds and the central dehydrated monovalent ion integral to formation and stabilization. Unimolecular (**b**) parallel and (**c**) antiparallel G-quadruplex topologies. Adapted from (1,6,10,11). Dark lines follow the nucleic acid strand, arrowheads denote strand directionality, and gray boxes denote quadruplexes. The examples drawn here are for sequences having three quartets.

plant stress wherein concentrations of K^+ and various osmolytes increase (see below).

RNA and DNA folding events have the potential to alter gene expression *in vivo* (19,20), and one such event is G-quadruplex formation. Using bioinformatics search methods such as the program 'Quadparser', ~375 000 putative G-Quartet Sequences (GQS) were identified in the human genome (21,22), where G-quadruplexes are found primarily in the intergenic regions and are enriched in telomeres (11,23). A large percentage (>40%) of human promoters contain at least one GQS sequence, implicating GQS in potential regulatory functions (24,25). In addition, several prokaryotes have also been shown to have a higher frequency of GQS motifs in upstream regulatory regions (26,27).

Genes containing GQS have been implicated in control of transcription, translation, stability and structure. Transcription of several oncogenes can be controlled by the promoter region GQS and binding of several ligands. A human c-*MYC* oncogene and a KRAS proto-oncogene are activated by protein binding to the GQS and subsequent destabilization of the structure (28–30). The same genes, as well as PDGF, can be inhibited by other binding factors (31–36). The potential for genome-wide response to regulatory region DNA GQS binding of ligands and protein factors has also been examined (37,38). Control of translation by GQS is evidenced by the ESR1 mRNA (39), the human *NRAS* proto-oncogene (40,41), the MT3 metalloproteinase mRNA (2), and *ZIC-1*RNA (42), all of which have repressed translation when a stable GQS forms in their 5′-untranslated regions (UTRs). G-quadruplex motifs have also been implicated in stability and post-transcription processing, for example in IGFII (Insulin-like growth factor II) mRNA, which has a GQS in its 3′-UTR (43). GQS can bind to proteins, such as the DNA thrombin aptamer (44–46) and FMRP (Fragile X Mental Retardation Protein), which binds to its own mRNA at a binding site that contains a GQS (47–49). In addition, an intermolecular G-quadruplex forms in HIV-1 genomic RNA dimers (50,51), and many GQS can bind small molecules, typically of a planar, aromatic nature (52,53).

Besides *Homo sapiens*, GQS have been reported for several other eukaryotic genomes, including *D*rosophila *melanogaster* and *Mus musculus* (54), however plant genomes have not been evaluated. Plants are of particular interest to GQS formation because under drought stress, K^+ ion concentrations in the cell can increase, which has the potential to drive G-quadruplex formation. For example, under drought or in high salinity soils, plants increase cytosolic K^+ concentrations up to 700 mM in order to avoid cellular dehydration (55,56).

The model plant *Arabidopsis thaliana*, with a completely sequenced genome and tractable genetics, lends itself well to investigation of G-quadruplex formation. Herein, we searched the genome of *A. thaliana* for GQS of various motifs and found that the prevalence and distribution of GQS in the genome varies according to GQS pattern. Functional analysis of genes with GQS suggests putative roles in modulating gene expression. Formation and relative stabilities of representative *Arabidopsis* GQS were verified experimentally using circular dichroism (CD) K^+ titrations and UV-detected thermal denaturation.

MATERIALS AND METHODS

Data sets searched

Genomic sequences were obtained from the TAIR9 version of the *Arabidopsis* genome, released 19 June 2009 (57). We searched for the presence of GQS in specific components of the genome, including 5′-UTRs, 3′-UTRs, coding sequences, introns, intergenic regions and upstream promoter sequences (defined as 1-kb upstream of the 5′-UTR or transcription start site). We utilized the major annotation provided by TAIR. Full chromosome sequences from the TAIR9 version of

the *Arabidopsis* genome were also searched. For comparisons, unmasked genomic sequences for 14 plant species, as well as *D. melanogaster* and *M. musculus* were evaluated. The following genomes were analyzed (see Supplementary Table S1 for details of common name, assembly, release data and server): *A. lyrata*, *Brachypodium distachyon* (58), *Drosophila melanogaster* (59), *Glycine max* (60), *Lotus japonicus* (61), *Manihot esculenta*, *Medicago truncatula* (62,63), *Mimulus guttatus*, *Mus musculus* (64), *Nicotiana tabacum* (65), *Oryza sativa ssp. indica* (66,67), *Oryza sativa ssp. japonica* (68,69), *Populus trichocarpa* (70), *Physcomitrella patens* (71), *Sorghum bicolor* (72), *Vitis vinifera* (73) and *Zea mays* (74). Genomic region sequences for *Oryza sativa ssp. japonica* (intergenic, coding sequences, exons, introns and UTRs) were also analyzed.

Programs used

Searches for G-quadruplex-forming sequences were performed using the program Quadparser (21), which scans input sequences for a given pattern of nucleotides. Several folding motifs were used, following the form $d(G_{X+}L_{1-N})_{3+}G_{X+}$ where 'X+' was defined as 2+ or 3+ (depending on the search), and L was 1, 1–2, 1–3, 1–4, or 1–7. These definitions correspond to G-quartets of varying stabilities, with larger values of X and smaller values of N generally corresponding to more stable sequences.

The GQS reported here are non-overlapping, such that any continuous stretch of GQS sequence will count as one, no matter the number of registers. In addition, we note that the choice of *Arabidopsis* data sets searched slightly affects the number of GQS found, which contributes to small differences in total GQS number. For example, searching different regions of the genome, such as intergenic regions, genic regions and coding sequences, separately produced 1187 GQS (Table 1). However, searching whole chromosome sequences yielded 1219 GQS. These small differences can be attributed to GQS that span the different regions of the genome. As such, both search methods were used for this study: chromosome searches enabled testing for significance (see below), while searching the genome regions allowed comparison of GQS across different components of the genome.

Table 1. Distribution of GQS motifs in the *Arabidopsis* genome

GQS Motif	Genome[a]	Intergenic[b]	Genic[c]	Coding[d]	Genic: intergenic
G_{3+} L_{1-7}	1187	827 (70%)	360 (30%)	263 (22%)	0.4
G_{3+} L_{1-3}	237	163 (69%)	74 (31%)	41 (17%)	0.4
G_{2+} L_{1-4}	43 117	8561 (20%)	34 556 (80%)	30 555 (71%)	4.0
G_{2+} L_{1-2}	12 340	1824 (15%)	10 516 (85%)	9415 (76%)	5.8
G_{2+} L_1	8188	901 (11%)	7287 (89%)	6633 (81%)	8.1

Numbers and percentages of GQS in different regions of the *Arabidopsis* genome. [a]Genome is comprised of [b]intergenic and [c]genic, while [d]coding is a subset of [c]genic and includes all gene models. Quadparser search parameters included G and C patterns to account for both sense and antisense strands.

Searches for G-quadruplex-forming sequences also included C-rich sequences, $d(C_{X+}L_{1-N})_{3+}C_{X+}$, which would form GQS in the complementary strand of the genomic DNA, although not in the transcript, or might form the i-motif in the C-rich strand (75). It should be mentioned that, as a result of the search parameters, some overlap in counted sequences occurred across GQS motifs: output sequences from less restrictive GQS criteria include sequences that follow stricter criteria; for example, sequences of the type $G_{3+}L_{1-3}$ occur within the $G_{3+}L_{1-7}$ and $G_{2+}L_{1-4}$ searches.

Functional analysis of the gene products was conducted using the BiNGO 2.3 plugin (76) for the Cytoscape 2.6.0 visualization program (77,78). BiNGO assesses over- or underrepresentation of genes in a user-defined category as compared to the entire *Arabidopsis* annotation. Here, the GQS-containing GO annotated genes were compared against the GO annotation for the entire *A. thaliana* transcriptome. A full GO analysis was performed for genic regions, using GO definitions as of 8 October 2008 and all of the GO terms: Molecular Function (MF), Biological Process (BP) and Cellular Component (CC) (79). Only loci with at least one GQS were included and no locus was included twice regardless of the number of GQS it contained. BiNGO uses the hypergeometrical statistical test, which is equivalent to an exact Fisher test, along with a Benjamini and Hochberg False Discovery Rate (FDR) correction at a significance level of 0.05.

To assess the significance of GQS patterns in *Arabidopsis*, we employed a windowed Markov model to generate simulated chromosomes. The mock chromosomes were generated in windows of fixed sizes, where nucleotide composition or di-nucleotide frequencies were modeled about corresponding windows along the corresponding real genomes. This corresponds to Bernoulli (or categorical random distribution) and Markov chains respectively. The process of generating a mock genome considers in turn every chromosome in the genome. For each chromosome, data are collected in a window of some fixed size, and used to generate a segment of quasi-random chromosome of the same length. The process is iteratively repeated one window length downstream until the end of the real chromosome is reached. We wrote the computer software using the C++ language. The programs are available for Mac OS X and Linux. For this instance, pattern hits were counted using the software grep available on Linux computers.

Oligonucleotide preparation

RNA oligonucleotides with the following sequences were purchased from Dharmacon (G-quartet forming stretches are underlined):

G_3L_{221}: 5′-GGGUCGGGUUGGGCGGG
G_3L_{444}: 5′-GGGUUUUGGGACAUGGGCUUGGGG
G_3L_{444}+FLANK: 5′-UGGGUCCUUUAAGUGUUUCUC CUAUGGGUUUUGGGACAUGGGCUUGGGGU
G_2L_{111}: 5′-GGAGGAGGAGGA
G_2L_{444}: 5′-GGAGCCGGAGUCGGAAUGGG

As an example of the shorthand notation adopted, G_3L_{221} indicates three G's interspersed by loops of 2, 2 and 1. These represent sequences from *Arabidopsis* genes as follows: G_3L_{221}: At1g07180 (*NDA1, non-phosphorylating NAD(P)H dehyrogenase*); G_3L_{444}: At5g53580 (MNC6.12, aldo/keto reductase family protein); G_2L_{111}: At2g39320 (T16B24.4, OTU-like cysteine protease family protein); G_2L_{444}: At1g44020 (F9C16.23, DC1 domain-containing protein). These sequences were chosen because they represented different types of candidate GQS motifs; in addition the G_3L_{444} sequence was chosen because of the strong flanking sequence that forms a predicted alternative base paired structure with a free energy of −15.6 kcal/mol, which was included in the oligonucleotide G_3L_{444}+FLANK. RNA oligonucleotides were first dialyzed against 100 mM LiCl, which does not support quadruplex formation, for 8 h to remove associated cations and then dialyzed against distilled and autoclaved water for 4 h to remove excess LiCl. Finally, RNA oligonucleotides were dialyzed overnight against 10 mM Li Cacodylate (pH 7.0). All dialysis was performed using a six well microdialysis apparatus from Gibco-BRL Life Technologies with a flow rate of 25 mL/min. To favor monomeric species, RNAs were renatured in the absence of potassium ions at 85°C for 1 min and then allowed to cool at room temperature before performing experiments. Native gels confirmed that RNA oligonucleotides of the above mentioned concentration and renaturation conditions form largely unimolecular structures (Supplementary Figure S1).

CD

CD spectroscopy was performed using a Jasco CD J810 Spectropolarimeter and analyzed with Jasco Spectra Manager Suite software. RNA samples were prepared as described above to a concentration ~5 μM. Spectra were acquired at 20°C over a wavelength range of 210–320 nm, with data collected every nanometer at a bandwidth of 1 nm. Reported spectra are an average of three scans at a response time of 4 s/nm. Data are buffer subtracted, normalized to provide molar residue ellipticity values, and smoothed over 5 nm (80). Molar residue ellipticity is reported in order to normalize for concentration differences and oligonucleotide length.

The amount of K^+ necessary to drive quadruplex formation was of interest. To determine $K^+_{1/2}$ values, ellipticity data (ε) were fit with KaleidaGraph v. 3.5 (Synergy software) according to the two-state Hill equation in which K^+ ions are taken up in the U to F transition:

$$\varepsilon = \varepsilon_F + \frac{\varepsilon_U - \varepsilon_F}{1+([K^+]/[K^+_{1/2}])^n} \quad (1)$$

where ε_F is the normalized CD signal corresponding to fully folded GQS, ε_U is the normalized CD signal for the unfolded GQS, $[K^+]$ is the potassium ion concentration, $K^+_{1/2}$ is the potassium ion concentration needed to fold half the RNA, and n is the Hill coefficient. Data are consistent with a two-state model ('Results' section).

UV thermal denaturation

RNA samples were prepared as described above at a concentration of ~5 μM in 10 mM Li Cacodylate (pH 7.0), with monovalent salts added before renaturation. Native gels confirmed formation of one monomeric species, as described earlier. Thermal denaturation experiments ('melts') were performed on a Gilford Response II spectrophotometer, with absorbances recorded every 0.5°C over the temperature range of 5–95°C and 95–5°C. Similar profiles were obtained for forward and reverse melts consistent with reversibility of the unfolding transition. Absorbance was detected either at 260 nm to observe the standard increase in A with temperature, or at 295 nm to observe the quadruplex-specific decrease in absorbance with temperature (a so-called 'inverse' melt) (81). Cuvettes with a 0.5-cm pathlength were used for 260 nm melts, while cuvettes with a 1-cm pathlength were used for 295 nm melts, owing to the smaller extinction coefficient at this wavelength. Data were fit with KaleidaGraph v. 3.5 (Synergy Software) using a Marquadt algorithm for non-linear curve fitting.

RESULTS

Prevalence and significance of GQS in *Arabidopsis*

We began our search for G-quadruplex sequences in the *Arabidopsis* genome using the standard definitions of a G-quadruplex forming sequence, $(G_{3+}L_{1-7})_{3+}G_{3+}$ and $(C_{3+}L_{1-7})_{3+}C_{3+}$ (21,22), referred to herein as 'G_3L_{1-7}' or simply 'G_3' and identified 1187 GQS (Table 1), which corresponds to a genomic density of 9.3 GQS/Mb (Table 2). Of the ~1200 GQS found in *Arabidopsis*, 329 have multiple registers, capable of forming more than one quadruplex structure. The importance of multiple registers is unclear, but it may contribute to overall stability, kinetics, and regulation.

Table 2. Density (D)[a] and Enrichment (E)[b] of GQS in various *Arabidopsis* genomic regions

GQS Motif	Genome	Intergenic		Genic		Coding	
	D (GQS/Mb)	D (GQS/Mb)	E	D (GQS/Mb)	E	D (GQS/Mb)	E
$G_{3+}L_{1-7}$	9.3	16.7	1.8[c]	4.6	0.5[c]	6.5	0.7
$G_{3+}L_{1-3}$	1.9	3.3	1.8	1.0	0.5	1.0	0.5
$G_{2+}L_{1-4}$	339.4	172.9	0.5	445.8	1.3	752.6	2.2
$G_{2+}L_{1-2}$	97.1	36.8	0.4	135.7	1.4	231.9	2.4
$G_{2+}L_1$	64.4	18.2	0.3	94.0	1.5	163.4	2.5

Provided are density and enrichment of GQS in different regions of the *Arabidopsis* genome from all gene models. Genome, intergenic, genic and coding are defined in Table 1.

[a]GQS density is defined as the total number of GQS per Megabase in the specified region; number of Megabases per region: whole genome 124.7 Mb, intergenic region 50.07 Mb, genic region 74.65 Mb, and coding sequence 39.59 Mb.

[b]Enrichment values are calculated as the GQS density of a region divided by the GQS density of the genome.

[c]These calculations are with $(G_3T_3A)_3G_3$ sequences (see text). As with Table 1, Quadparser search parameters included G and C patterns to account for both sense and antisense strands.

In comparison to the eukaryotes *H. sapiens*, *D. melanogaster* and *M. musculsus*, which have GQS densities of 115 GQS/Mb (24,25), 88.7 GQS/Mb (54) and 189.4 GQS/Mb, respectively, *Arabidopsis* has a lower number and density of $G_3L_{1–7}$ GQS (Figure 2). We also compared $G_3L_{1–7}$ GQS density in *Arabidopsis* to GQS density in 14 different plant species having a range of phylogenetic distances from *Arabidopsis*, including eudicot, monocot and moss species. We find that the GQS density of *A. thaliana* is similar to those of other dicot species, especially *A. lyrata* and *M. esculenta*, and to the moss *P. patens*. Monocots, such as *Z. mays*, *O. sativa*, and *B. distachyon* have significantly higher GQS densities (Figure 2, Supplementary Figure S2 and Table S1). A comparison of GC content and GQS density across the different plant genomes shows that GQS levels correlate reasonably well with the GC content of each genome (Supplementary Figure S2, green symbols, $R^2 = 0.76$). However, the same correlation is not true for all genomes. In particular, *H. sapiens* and *D. melanogaster* follow the same trend as plant genomes, but *M. musculus* does not.

Tandem repeats of two guanines (G_2) have also been reported to form G-quartet structures, albeit in a less stable form. For example, Hartig and coworkers studied differences in RNA GQS with G_2 and G_3 sequences and showed that G_2 sequences formed GQS but had melting temperatures about 25°C lower than equivalent G_3 sequences; in addition, all G_2 sequences had weaker CD signatures than G_3 sequences (82). Nonetheless, G_2 sequences can form G-quartets that can be quite stable, especially in the presence of high salt concentrations (82). The thrombin DNA aptamer, with the sequence $G_2T_2G_2TGTG_2T_2G_2$ is an example of a well characterized DNA G_2 quartet (44–46). We therefore also evaluated prevalence of GQS for stretches of two or more guanines and a maximum loop size of four ($G_2L_{1–4}$). This simplification of the search motif dramatically increased the number of hits, from ~1200 to ~43 000 (Table 1), providing a genomic density of 339 GQS/Mb (Table 2).

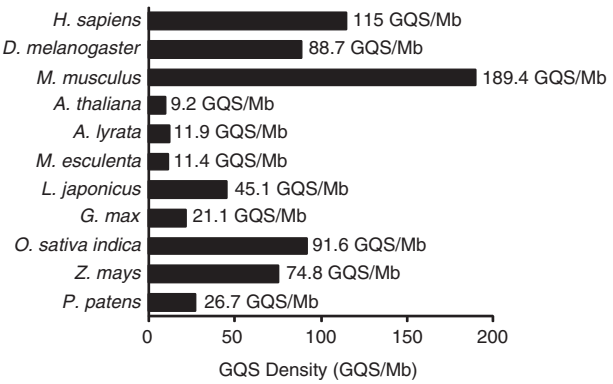

Figure 2. Density of GQS in different organisms. Density of GQS in each genome is represented by a bar and given in GQS/Mb. GQS here is defined as $G_3L_{1–7}$. GQS density is provided for each organism at the right-hand end of the bar. All numbers are from this study except *H. sapiens* which is from Huppert *et al.* (21) and Todd *et al.* (22). Common names: *Homo sapiens* (Human), *Drosophila melanogaster* (fruitfly), *Mus musculus* (mouse), *Arabidopsis thaliana* (mouse ear cress), *Arabidopsis lyrata* (lyrate rock cress), *Manihot esculenta* (cassava), *Lotus japonicus* (*Lotus japonicus*), *Glycine max* (soybean), *Oryza sativa indica* (rice—indica), *Zea mays* (corn) and *Physcomitrella patens* (*Physcomitrella patens,* a moss).

To assess significance of the number of GQS found in *Arabidopsis*, we applied a windowed Markov Model, maintaining dyad frequency, to generate randomized genomes, similar to the method of Huppert and Balasubramanian (21). Results of various window sizes are shown in Table 3. The AT patterns were used as a control to ensure that the generated random genome was a good representation of the real genome. Markov window sizes that are too small do not sufficiently randomize the genome, producing more hits for both the AT pattern control and the GC pattern of interest. On the other hand, larger window sizes yield smaller numbers of patterns, as GC-rich regions become diluted in the window. A Markov window of 100 was determined to be an appropriate mimic of the real genome, as the number of AT patterns (147 451) matched most closely that of the real genome (147 266) (Table 3). We find that GC patterns are underrepresented by a factor of 2.3 in the real genome when compared to the random genome (2875 expected, 1232 found). These sequences are even more underrepresented than in the human genome, where G_3 GQS are only underrepresented by a factor of 1.4 (21).

The G_2 patterns were examined using the Markov Model as well. Again, a window size of 100 was optimal, however, the G2 pattern was not underrepresented as the G_3 patterns were, with 44 168 G2 patterns expected, and 43 117 found. Nonetheless, G2 patterns may play important biological roles, as revealed by our functional analysis and drought stress analysis on these motifs, described below. Next, we investigated where these GQS sequences are located within the genome.

Table 3. Number of patterns in *Arabidopsis* and Markov simulated genome

	Window	X_3 $L_{1–7}$ *GC*	X_3 $L_{1–7}$ *AT*	X_2 $L_{1–4}$ *GC*	X_2 $L_{1–4}$ *AT*
Real		1232	147 266	43 117	746 324
Markov	50	5838	195 259	63 307	816 536
Markov	75	3776	165 195	50 827	771 304
Markov	**100**	**2875**	**147 451**	**44 168**	**743 265**
Markov	150	1977	125 727	36 554	708 312
Markov	200	1509	113 870	32 461	687 647
Markov	400	847	91 457	25 208	646 856
Markov	1000	421	73 006	19 076	608 637
Markov	2000	282	64 350	16 112	588 057
Markov	4000	191	58 097	14 113	572 833

Number of GC and AT patterns in the real *Arabidopsis* genome and a windowed Markov model simulated genome. See 'Materials and Methods' section for more details. The window size that accurately simulated the AT pattern of the *Arabidopsis* genome is 100 and is in bold text.

Location of GQS in the *Arabidopsis* genome

We found that the distribution of GQS throughout the genome is not uniform. First, the prevalence of G_3L_{1-7} GQS in genic and intergenic regions of the genome was compared. As described earlier, the density of G_3L_{1-7} GQS in the *Arabidopsis* genome is 9.3 GQS/Mb (Table 2). Seventy percent of these G_3L_{1-7} GQS are present in the intergenic regions, yielding an enrichment value (intergenic density/whole genome density) of 1.8. It follows that G_3L_{1-7} GQS are depleted in the genic regions, with a lower GQS density than the remainder of the genome. Here, the corresponding values are 30% (Table 1) and 0.5. (Table 2), and the ratio of DNA G_3L_{1-7} GQS in the genic to intergenic regions is 0.4 (Table 1, right-most column).

As mentioned earlier, we found a correlation between genomic GQS density and GC content for 15 plant species and a few of the non-plant eukaryotes. However, this correlation did not extend to sub-regions of the *Arabidopsis* genome. When genic and intergenic regions of the genome are considered separately, the G_3 GQS densities are opposite of the GC content. The genic region has a GC content of 38.9% and the intergenic region has a GC content of 31.1%, while the genic region has a G_3 GQS density of 4.6 GQS/Mb, much lower than the intergenic region value of 16.7 GQS/Mb.

We examined the genomic regions of the well-annotated monocot *O. sativa ssp. japonica* to see if this trend extended to other plant species. In *O. sativa*, even though the GC content and GQS density is higher than in *Arabidopsis*, we observe the same inverse correlation between the genic and intergenic regions. The genic region has a GC content of 45% and a GQS density of 82.6 GQS/Mb, while the intergenic region has a lower GC content of 41.5% but a higher GQS density of 127.9 GQS/Mb.

Approximately 160 (20%) of the intergenic GQS were found to correspond to the *Arabidopsis* telomeric sequence, $(G_3T_3A)_{3+}G_3$. These sequences were found to be located not only at chromosome ends but also in interstitial sites near the centromeric regions. This positioning of the telomeric sequence has been previously explained by chromosomal rearrangements, such as the evolutionary combination of two chromosomes, Robertson fusions, or arm inversions (83–85). To determine if the large numbers of telomeric sequences accounted for the GQS enrichment in the intergenic region, these sequences were removed from the GQS density calculation; nonetheless, the intergenic region was still enriched in GQS, with an enrichment value of 1.6 relative to the whole genome, similar to the value of 1.8 presented above. Since the effect of telomeric sequences on GQS density and enrichment values was not large, we used all GQS sequences for subsequent comparisons and calculations.

Because we were interested in different GQS motifs and relative stabilities of the *Arabidopsis* GQS, we next tallied GQS statistics considering different loop lengths and numbers of guanines. In particular, we looked at whether certain loop lengths are more common than others in the *Arabidopsis* genome. First, we confined the loop to a maximum of three nucleotides, G_3L_{1-3}, which decreased the number of GQS found for the entire genome to just 237 (Table 1). These sequences, which should be very stable, are infrequent in the genome, with only 1.9 sequences per Mb (Table 2). Even though there are almost 4-fold fewer unique GQS than for G_3L_{1-7}, the relative enrichments, 1.8 for the intergenic region and 0.5 for the genic region, are the same as for G_3L_{1-7} sequences (Table 2).

Next, we changed the GQS definition to G_2L_{1-4}, which includes stretches of two or more guanines and a maximum loop size of four in the DNA region. As described, expanding the definition increased the number of GQS found to ~43 000 (Table 1), corresponding to a GQS density of 339 GQS/Mb (Table 2). Remarkably, upon decreasing the number of G's in the search motif, the density of GQS shifted from favoring intergenic to favoring the genic region (enrichment value of 1.3 in G_2L_{1-4}) (Table 2). In fact, whereas just 30% of G_3L_{1-7} sequences are predicted to reside in genes, the vast majority of G_2L_{1-4} sequences (80%) are predicted to be genic (Table 1) and of those, nearly all (88%, or 30 555/34 556) are located in coding sequences. Apparently, location in the genome depends on GQS type ('Discussion' section). Unlike the negative correlation between GQS density and GC content with G_3 sequences, GQS density and GC content have a positive correlation for G_2 sequences.

Next, we limited the size of the loop of the G_2 motif to a maximum of two. As expected, this led to the identification of fewer GQS, with 12 340 sequences found (Table 1). While the number of GQS decreased, distribution in the genome was affected only slightly: there was a modest increase in percentages for genic (80 to 85%) and coding sequences (71 to 76%) (Table 1). Lastly, we constrained the loop size of the G_2 motif to just one nucleotide. As expected, this further decreased the number of GQS in the genome, to 8188 (Table 1), but it maintained approximately the same GQS distribution across the genome, with further small increases in the percentage of genic (85 to 89%) and coding (76 to 81%) sequences. With this definition, 91% (6633 out of 7287) of the GQS were located in coding sequences. The prevalence of G_2 sequences with short loops in genic and coding sequences is also reflected in enrichment values (Table 2): in going from L_{1-4} to L_{1-2} to L_1, enrichments increased from 1.3 to 1.4 to 1.5 for the genic region and from 2.2 to 2.4 to 2.5 for the coding sequence. Moreover, the ratio of DNA GQS in genic to intergenic ratios increased from 0.4 to 8.1 as the number of G's was decreased and the loop was shortened (Table 1, right-most column). Thus, the G_2 motif is enriched in genic regions, especially the coding sequences, and this enrichment is further enhanced for GQS with shorter, and somewhat more stable, loops.

We were especially interested in whether the shorter GQS might be more prevalent in RNA as compared to DNA. Comparison of genic RNA GQS to intergenic DNA GQS was therefore made (Figure 3). In this accounting, 'RNA' is defined as genic, which includes coding sequences, UTRs and introns, but only G-rich sequences, while 'DNA' is defined as intergenic and includes

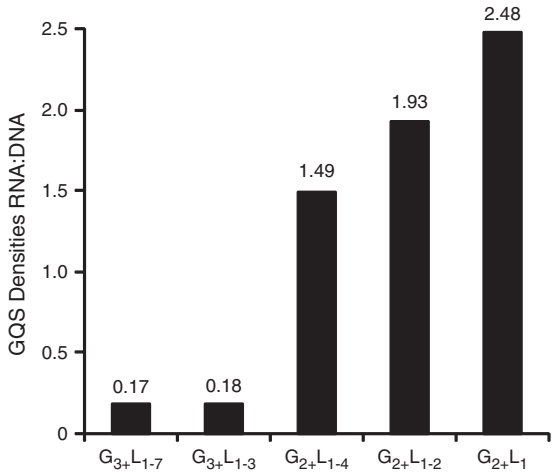

Figure 3. Ratio of RNA GQS density to non-genic DNA GQS density for different GQS motifs. DNA GQS density includes G and C sequences in the intergenic region only (Table 2, column 3), since this will not lead to GQS in RNA. RNA GQS density includes only the G sequences found in the genic regions (Table 5, column 3). For example, $G_{3+}L_{1-7}$, RNA GQS density is 2.9 and the DNA intergenic region GQS density is 16.7, leading to a ratio of 0.17.

both G- and C-rich sequences. The RNA:DNA GQS density ratio highlights the differences in genomic distribution between G_3 and G_2 GQS definitions. The G_3 sequences are found mostly in intergenic regions and rarely in RNA, with RNA:DNA GQS density values of 0.17 and 0.18 for loops of 1–7 and 1–3, respectively (Figure 3). In contrast, G_2 sequences are more prevalent in RNA than intergenic regions, with RNA:DNA density ratios of 1.5 or greater; this holds despite the fact that both DNA strands (i.e. G and C patterns) are included in the accounting for the intergenic region. We also note that the RNA:DNA density ratio increases within the G_2 motif as L decreases from $L_{1–4}$ to $L_{1–2}$ to L_1, with values of 1.5, 1.9 and 2.5, respectively (Figure 3).

Characterization of GQS found in the intergenic region

Recent studies on *Arabidopsis* have revealed that some RNAs are transcribed from intergenic regions. From whole genome tiling arrays, it has been found that ~19–23% of the *Arabidopsis* intergenic region is transcribed, in so-called 'intergenic transcriptional units' (86,87). We therefore determined where the intergenic GQS occur in relation to these transcriptional units (TU). A summary can be found in Table 4. For $G_3L_{1–7}$ GQS, of the 827 intergenic GQS (Table 1), only 22 (3%) fall in an intergenic TU, corresponding to a GQS density of just 2.3 GQS/Mb. This leaves a large non-transcribed intergenic GQS density of 20 GQS/Mb (Table 4). In other words, the density of $G_3L_{1–7}$ GQS is a striking 8.7-fold higher in non-transcribed versus TU intergenic regions. Moreover, the GQS density of the intergenic TU of 2.3 is 2-fold less than the genic region density of 4.6 per Mb (Table 2).

Regarding the $G_2L_{1–4}$ GQS, a larger number of these motifs (8561) are found in the intergenic region (Table 1) compared to the $G_3L_{1–7}$, as expected, with 686 (8%) overlapping with transcriptional units (Table 4). This corresponds to an intergenic region TU GQS density of 72 GQS/Mb. The remaining 7875 $G_2L_{1–4}$ GQS (92%) are located in the non-transcribed intergenic region, with a density of 194 GQS/Mb. Thus, the density of $G_2L_{1–4}$ GQS is also higher in non-transcribed versus TU intergenic regions, albeit not as much as for the $G_3L_{1–7}$ motif, with density ratios of 2.7- and 8.7-fold, respectively. These ratios suggest that GQS motifs may play a role in repressing transcription in intergenic regions ('Discussion' section).

Table 4. Distribution and Density (D) of GQS motifs in *Arabidopsis* intergenic regions

GQS Motif	Intergenic[a]	Transcribed units[b]			Non-transcribed units[c]			D non-TU/ D TU
		GQS	*D*[d]	%[e]	*GQS*	*D*[d]	%[e]	
G_3 $L_{1–7}$	827	22	2.3	3	805	20	97	8.7
G_2 $L_{1–4}$	8561	686	72.0	8	7875	194	92	2.7

Provided are distribution and density of GQS in transcribed (TU) and non-transcribed (non-TU) regions of the intergenic region.
[a]Intergenic region is comprised of [b]transcribed units and [c]non-transcribed units. Raw numbers (*GQS*) are provided.
[d]GQS density (*D*) is defined as the total number of GQS per Megabase in the specified region.
[e]Percentages were calculated relative to the intergenic region. Number of Megabases per region: intergenic region 50.070 Mb, TU 9.53 Mb, non-TU 40.54 Mb. Quadparser search parameters were set to include both G- and C-patterns.

Next, we examined whether GQS are preferentially localized to promoters. In the human genome, GQS are localized to promoter regions, with a 6-fold enrichment value compared to total genomic DNA, and at least one GQS present in 42.7% of promoters (24). We therefore assessed whether similar trends hold for *Arabidopsis*. In contrast, *Arabidopsis* $G_3L_{1–7}$ GQS are not over-represented upstream of genes. For example, only 317 GQS were found in promoter regions, corresponding to a density of just 9.5 GQS/Mb (using 33.20 Mb as the total length of all promoter regions), lower than the overall *Arabidopsis* intergenic region GQS density of 16.7 GQS/Mb (Table 2), and much less than the human promoter GQS density of 770 GQS/Mb (24). $G_2L_{1–4}$ GQS are more prevalent than $G_3L_{1–7}$ in promoter regions, with 8306 $G_2L_{1–4}$ GQS, corresponding to a density of 250 GQS/Mb, which is somewhat greater than the overall intergenic region density of 173 GQS/Mb (Table 2). Approximately 20% (8620) of all genes have at least one of these two types of GQS in the promoter region, suggesting a possible role of G_2 GQS in regulating via the promoter.

Characterization of GQS found in the genic region

As motivation for characterizing GQS in the genic region, we first asked, for any given GQS definition, how many genes or loci contain at least one GQS in the corresponding mRNA. [Gene models are named uniquely within a given open reading frame (ORF), thus a given locus can have more than one associated gene model, e.g. if there are alternatively spliced variants for a gene].

Of the 39 640 gene models in *Arabidopsis*, 215 gene models were found to contain ≥1 G_3 GQS in their RNA, which represents 0.5% of the total number of genes in *Arabidopsis* (Supplementary Table S2). In contrast, ≥1 $G_{2+}L_{1-4}$ GQS were found in about one-third of all gene models (33%) and loci (31%), suggesting the potential for widespread formation (Supplementary Table S2). A complete list of gene models that contain a GQS can be found in Supplementary Table S3.

Regulation of translation by G_3 GQS in RNA transcripts has been demonstrated in *Escherichia coli* and eukaryotic cells (2,82,88). Given the large number of G_2 quartets in *Arabidopsis* sequences corresponding to mRNAs, the location of these sequences within the genic region was determined. To examine GQS that would appear in the RNA, we limited searching to G patterns from the genic regions (i.e. C patterns were excluded), and divided results between coding sequences (cds), 5′- and 3′-UTRs and introns (Table 5). In the *Arabidopsis* genome, the majority (90%) of G_2 RNA GQS are in the cds, which has an ∼4-fold higher density than the non-coding UTRs (443, 110 and 105.5, for cds, 5′-UTR and 3′-UTR, respectively, Table 5). Within the cds, GQS are located towards the 5′-end (Supplementary Figure S3). For the G_3 sequences, the cds has a 2-fold higher GQS density than the non-coding UTRs (4.3, 2.6 and 2.2 for cds, 5′-UTR and 3′-UTR, respectively). Lastly, the intronic regions for both G_2 and G_3 motifs have much lower GQS density, with a coding/intronic density ratio of 13 for G_2L_{1-4} and 4.3 for G_3L_{1-7} (Table 5).

Potential biological functions of GQS in *Arabidopsis*

Because cellular K^+ concentrations often increase under drought stress, we considered if any of the genes containing G_2 GQS are differentially expressed when exposed to drought stress conditions. We used previously published tilling array data by Matsui *et al*. (86) that reported drought-responsive loci. As mentioned earlier, 31% of all loci in *Arabidopsis* contain at least one G_2L_{1-4} GQS (10 382/33 518). Matsui *et al*. (86) report that 5508 loci are drought responsive, which corresponds to ∼16% of all loci in *Arabidopsis*. Of those loci, 45% (2474) have at least one GQS. By chi-square (χ^2) analysis of these values—31% of all loci have a GQS versus 45% of all drought-responsive loci have a GQS—we determined that drought-regulated loci are indeed significantly more likely to have a GQS than when considering all loci, ($P < 0.0001$). The reverse analysis, inquiring if loci that have a GQS are more likely to be drought responsive than loci without GQS, is also true—16% of all loci are drought responsive versus 24% of GQS loci are drought responsive ($P < 0.0001$).

Given the enrichment of G_2 sequences in the genic portion of the genome, we next investigated whether these sequences are overrepresented in certain functional classes of loci. The functions of gene products encoded by genes containing at least one GQS were examined using Gene Ontology (GO) codes as analyzed with the program BiNGO (76). A number of GO terms were found to be overrepresented in proteins encoded by G_2-containing genes. A sample of unique, over- and underrepresented GO terms for the $G_{2+}L_{1-4}$ GQS definition and the respective *P*-values, all <1E-8, are provided in Table 6, and a complete list of over- and underrepresented GO terms can be found in Supplementary Table S4. In addition, GO analysis was performed with other G_2 GQS motifs, providing over- and underrepresented GO terms with less statistical significance owing to the smaller sample sizes (Supplementary Tables S5, S6, and S7).

The GO term with the greatest statistical significance for the G_2L_{1-4} motif is 'catalytic activity', with an exceptionally low *P*-value of 9E-65 (Table 6). This term corresponds to genes that code for enzymes. According to this analysis, 45% of 'catalytic activity'–annotated loci have one or more G_2L_{1-4} GQS, which is much larger statistically than the 33% of all loci that have such a GQS. Other highly significant GO terms include nucleotide binding and multicellular organismal development, as well as specific catalytic activities such as post-translational protein modification, kinase activity, transferase activity and helicase activity, all of which have $P < 7E\text{-}17$. One possibility is that during stress in *Arabidopsis* G-quartet structure formation is enhanced, which represses expression of these genes allowing metabolism to decrease ('Discussion' section).

Underrepresented GO terms were also determined with BiNGO. Interestingly, genes in a number of GO categories

Table 5. Distribution and Density (D) of GQS motifs in *Arabidopsis* genic RNA

GQS Motif	Genic[a]		Coding[b]			5′-UTR[c]			3′-UTR[d]			Intron[e]		
	GQS	D^f	*GQS*	D^f	%[g]	*GQS*	D^f	%[g]	*GQS*	D^f	%[g]	*GQS*	D^f	%[g]
$G_{3+}L_{1-7}$	225	2.9	174	4.3	77	10	2.6	4	14	2.2	6	27	1.0	12
$G_{3+}L_{1-3}$	43	0.6	28	0.7	65	4	1.1	9	4	0.6	9	7	0.3	16
$G_{2+}L_{1-4}$	19 985	257.8	17 989	443.1	90	417	110.0	2	657	105.5	3	922	34.3	5
$G_{2+}L_{1-2}$	5435	71.1	4897	120.6	90	124	32.7	2	128	20.5	3	286	10.6	5
$G_{2+}L_1$	3496	45.1	3174	78.2	91	74	19.5	2	63	10.1	2	185	6.9	5

Provided are distribution and density of GQS in different regions of the genes from all gene models.
[a]Genic region is comprised of [b]coding, [c]5′-UTR, [d]3′-UTR and [e]Intron. [f]GQS density is defined as the total number of GQS per Megabase in the specified region. Number of megabases per region: genic region 74.65 Mb, CDS 39.59 Mb, 5′-UTR 3.62 Mb, 3′UTR 6.02 Mb and intron 25.43 Mb. [g]Percentages were calculated relative to the genic region. Raw numbers (*GQS*), GQS densities (*D*) and percentages (*%*) of GQS. Quadparser search parameters were set to include only G-patterns, which will be found in RNA, and exclude C-patterns.

Table 6. Functional analysis of genes with at least one G_2L_{1-4} GQS present in the RNA

GO ID[a]	GO Cat[b]	GO term[c]	GQS genes[d]	All genes[e]	% GQS genes[f]	*P*-value[g]
Overrepresented						
0003824	MF	Catalytic activity	2894[h]	6393[i]	45	9E-65
0006468	BP	Protein amino acid phosphorylation	506	798	63	1E-53
0016740	MF	Transferase activity	1118	2176	51	2E-49
0016301	MF	Kinase activity	661	1151	57	2E-48
0043687	BP	Post-translational protein modification.	579	985	59	2E-46
0005478	MF	Transporter activity	504	993	51	1E-19
0000166	MF	Nucleotide binding	490	980	50	2E-17
0048856	BP	Anatomical structure development	359	681	53	5E-17
0007275	BP	Multicellular organismal development	441	871	51	7E-17
0016020	CC	Membrane	1011	2266	45	3E-16
0022414	BP	Reproductive process	275	502	55	8E-16
Underrepresented						
0000496	MF	Base pairing	0	631	0	<1E-99
0006412	BP	Translation	174	1129	15	4E-51
0010467	BP	Gene expression	293	1487	20	1E-41
0003723	MF	RNA binding	179	983	18	7E-30
0000154	BP	rRNA modification	1	70	0.01	3E-9

Provided are overrepresented and underrepresented gene ontology[a] (GO) ID numbers, [b]GO categories (Cat) and [c]GO term for gene products encoded by pre-mRNA with at least one $G_{2+}L_{1-4}$ GQS. Included are [d]the number of genes (scored if GQS is in CDS, 5′-UTR, 3′-UTR, or introns) with a GQS that are annotated for the listed GO term, and [e]the total number of genes in *Arabidopsis* with the listed GO term. Also included are [f]the percentage of genes with GQS with a given GO term and [g]the appropriate *P*-value, as determined using the BiNGO program. [h]The total number of GO-annotated genes with a GQS in G_2L_{1-4} is 9097. [i]The total number of GO-annotated genes in *A. thaliana* is 25 179. Table is sorted in order of increasing *P*-value. Some GO terms are sub-categories of others. Complete list is provided in Supporting Information Supplementary Table S2.

were found to have very significant depletion in G_2L_{1-4} GQS. In particular, genes with GO terms for base pairing, translation, and rRNA modification had *P*-values for underrepresentation of <1E-99, 4.2E-51 and 3.0E-9, respectively; in the case of base pairing, none of the 631 genes had a GQS. The genes in these activities correspond to non-coding RNAs such as tRNAs, rRNAs and snoRNAs whose function depends on specific RNA secondary or tertiary structures. We hypothesize that GQS are underrepresented in these RNAs because G-quartet structures would disrupt base pairing and therefore function ('Discussion' section).

About 14% of the genes in *Arabidopsis*, or 4626, are known to have alternative splice variants (89). Environmental conditions and stresses (90,91), developmental stages (92) or tissue localization can favor the production of alternatively spliced transcripts (93). We searched the annotated alternative splice variants for the presence of GQS and found that over 8000 gene models (this number counts all the individual splice variants) have at least one G_2 GQS. Moreover, we found that 108 genes have at least one splice variant that contains a GQS and one that either does not have a GQS or has one located in a different genic region. For example, CTC1 (Conserved Telomere Maintenance Component 1) has two splice variants, one without a GQS, and one with a G_2 GQS located in intron 15. Another example is SPL10, which has four splice variants: one which does not contain a GQS, one which has a GQS in its 5′-UTR and two which have a GQS in intron 1. Although we have not found any readily apparent correlation between GQS location and alternative splicing, the possibility remains that GQS play a functional role in the alternative splicing or expression of some genes.

Experimental evidence for G-quartet formation

To verify that the GQS identified using bioinformatics methods actually form quadruplex structures, folding and thermal denaturation experiments were performed on select RNA sequences *in vitro*. We chose short RNA oligonucleotides that correspond to specific GQS folding motifs that are present in specific *Arabidopsis* genes. The unimolecular nature of the folding transition and the ability to form a G-quadruplex were assessed by native PAGE (Supplementary Figure S1). As shown in Supplementary Figure S1A, all representative oligonucleotides ran primarily as single bands and in the order expected, with the exception of G_2L_{111} (see below) and G_3L_{444}+FLANK (last lane), which likely has a contribution from a fold involving the flanking nucleotides. Also, the ability of a G2 motif to form a G-quadruplex structure is confirmed in Supplementary Figure S1B, where adding an increasing number of repeats leads to a faster mobility species once all four repeats are reached (lane 4), which is further confirmed by an oligonucleotide of the same length but a G to A single mutation that migrates with normal mobility (lane 5). Quadruplex formation is further confirmed by CD spectra and UV-melts (see below).

CD-detected K^+ titrations were used to judge whether a quadruplex formed and to determine ion affinity, and UV-detected thermal denaturations were used to assess quadruplex thermal stability. G-quartets have a unique CD signature that depends on topology: parallel GQS have a positive peak at 265 nm and a negative peak at 240 nm, while antiparallel GQS have a positive peak at 295 nm and a negative peak at 260 nm (1,7).

Sequences chosen and their associated gene IDs are provided in 'Materials and Methods' section, and relevant thermodynamic parameters are provided in Table 7. CD spectra for sequences representative of specific folding motifs of interest are provided in Figure 4a, normalized to concentration and oligonucleotide length (94). Height of the normalized CD peak gives an indication of population. Peak heights in Figure 4a are in the order G_3L_{221}>G_2L_{111}>G_3L_{444}>G_3L_{444}+FLANK>G_2L_{444}. A native flanking sequence of 25 5′-nt was added to the G_3L_{444} oligo (herein notated G_3L_{444} + FLANK) to provide a more biological context and to allow for the possibility of competing secondary structure. This flanking sequence interacted with the nucleotides that would otherwise form the GQS to give a predicted (95) free energy of −15.7 kcal/mol. Almost all of the sequences tested appear to have a fully parallel GQS structure, except for G_2L_{111} which has both parallel and antiparallel character, exemplified by the shoulder in the CD spectrum at 290–300 nm. This requires further analysis, and evidence for another structure can be seen in the native gels (Supplementary Figure S1).

To verify GQS formation and assess thermal stability, we performed UV melts on these RNAs in the background of 100 or 1000 mM Li^+, Na^+, or K^+. Data were acquired at 260 nm, as well as at 295 nm, which gives a quartet-specific inverse melt in which absorbance decreases with temperature which is associated with unfolding of a G-quadruplex (81) (see Figure 4c and d for sample melts). (Inverse melts were observed for G2 sequences as well (data not shown), which along with the data in Supplementary Figure S1B supports their ability to form quadruplexes.) All RNA GQS were most stable in K^+, followed by Na^+ and then Li^+, as expected for GQS unfolding. The strong K^+ preference of these structures is also illustrated by the drastic increase in T_m of 25 to 40°C in the presence of K^+ versus Na^+ (Table 7). In addition, G_2L_{444} only had a well-defined folding transition in K^+ but not Li^+ or Na^+. Melting temperatures of G_3L_{221} and G_2L_{111} oligonucleotides in K^+ were especially high, ≥85°C (Table 7). Overall, these data show at least partial formation of GQS at 100 mM K^+ ion concentration (see 'Discussion').

Table 7. Experimental values for G-quartet formation in *Arabidopsis* RNA

GQS Motif	Gene ID[a]	$K^+_{1/2}$ (mM)[b]	T_m (°C)[c]		
			Li^+	Na^+	K^+
G_3 L_{221}	At1g07180	8.0 ± 0.7	54	62	>85
G_3 L_{444}	At5g53580	42 ± 5	30	45	74
G_3 L_{444} + FLANK[d]	At5g53580	220 ± 70	ND[e]	ND	ND
G_2 L_{111}	At2g39320	30 ± 2	32	43	>85
G_2 L_{444}[d]	At1g44020	316 ± 1	ND	ND	62

Provided are thermodynamic parameters for G-quartet formation for RNA oligonucleotides from representative *Arabidopsis* genes. See 'Materials and Methods' section for full sequences.
[a]Gene ID identifies the particular gene from the *Arabidopsis* genome and sequences are provided in 'Materials and Methods' section.
[b]$K^+_{1/2}$ (K^+ concentration needed to fold half the RNA) values were determined by CD titrations and using Equation (1).
[c]T_m (melting temperature) values were determined by UV thermal melts at 100 mM monovalent salt concentration using the chloride salt.
[d]Melts for these oligonucleotides were performed at 1M salt concentration owing to their higher $K^+_{1/2}$ values.
[e]ND indicates that the T_m value could not be determined due to absence of a well defined folding transition.

DISCUSSION

In this study, GQS in the *Arabidopsis* genome were tallied by computational approaches and characterized by *in vitro* experiments. We categorized G_3 and G_2 sequences with different loop lengths. The G_3 sequences were more common in intergenic than genic regions, especially in the non-transcribed intergenic units, while the G_2 sequences were common in genic regions. Overall, G_3 sequences in *Arabidopsis* were even less common than in the human genome, especially in the promoter region, perhaps because the higher K^+ concentrations present during stress in plants provide a negative selective pressure. Within the genic regions, RNA GQS were mostly in the coding region, with one-third of all *Arabidopsis* genes containing at least one G_2 motif. Remarkably, these RNAs were overrepresented in certain classes of genes, especially those with catalytic activity, and underrepresented in other classes of genes, especially those that require base pairing for their function. Introns also had low GQS density. In addition, genes that are differentially regulated by drought stress are significantly more likely to contain a GQS than the genome as a whole. Experiments on representative RNAs from *Arabidopsis* confirmed that they form G-quartets under physiological conditions.

Formation and function of GQS in DNA and RNA

Formation of GQS in DNA was scored with both G and C patterns, allowing for putative formation of G-quartets in the sense and antisense strands of the genomic DNA. This was done because *in vitro* studies with self-complementary oligonucleotides have shown that quadruplex structures can exist in equilibrium with the DNA duplex (96) and that the i-motif in the C-rich strand might be favored by macromolecular crowding (75). In addition, during transcription, a single-stranded transcription bubble of 7–12 nt is formed, breaking the duplex structure and allowing for easier formation of a quadruplex in either strand of DNA (97). These DNA GQS thus have the potential for regulation at the level of transcription (98).

In addition to DNA, G-quartet sequences can form in RNA, where they have the potential for transcriptional, translational, or mRNA stability regulation (28–30, 39–43). The RNA quadruplex structure is expected to exist in equilibrium with alternative structures such as hairpins involving flanking sequence, which can control GQS stability. Equilibrium between the quadruplex and base-paired structures could be controlled by cellular conditions such as K^+ concentration (16,18).

We report a correlation between GC content and GQS density for 15 plant species, including monocots, dicots and a moss (Supplementary Figure S2). However, the

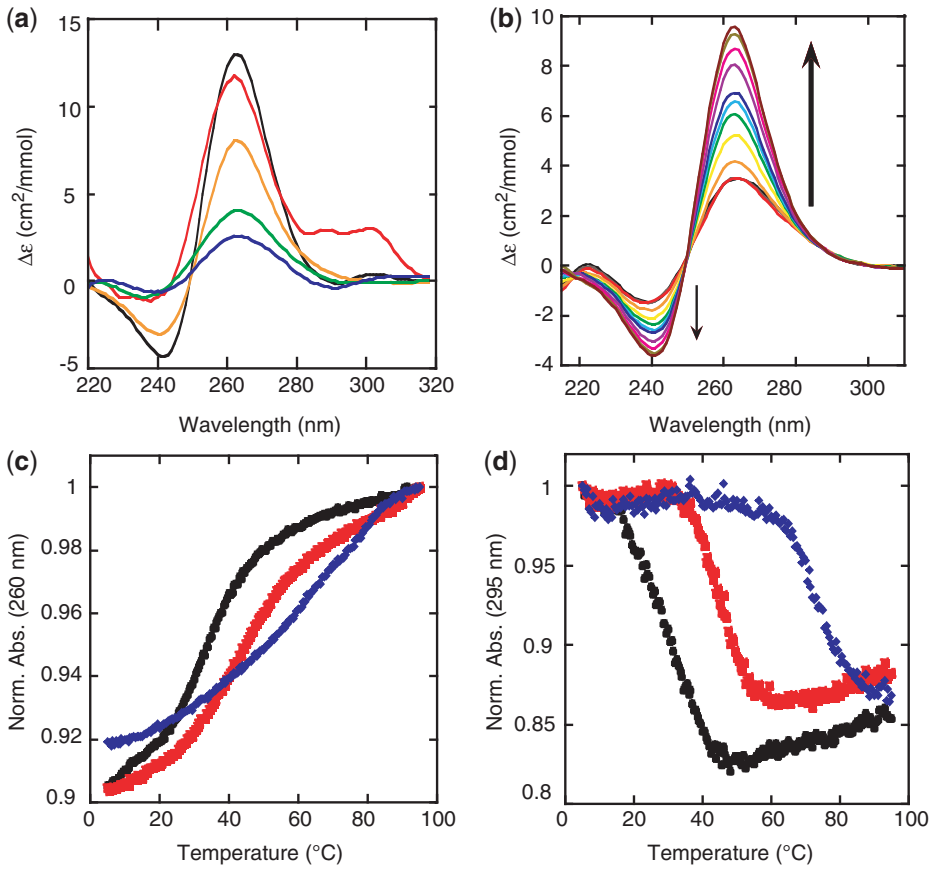

Figure 4. Formation of GQS RNA oligonucleotides (**a**) CD spectra of RNA GQS in 10 mM LiCacodylate (pH 7.0) and 150 mM KCl at 20°C: G_3L_{221} (black), G_2L_{111} (red), G_3L_{444} (gold), G_3L_{444} + FLANK (green) and G_2L_{444} (blue). The positive peak at 260 nm and the negative peak at 240 nm suggest that the RNA GQS adopt a parallel conformation. The G_2L_{111} RNA most likely has some antiparallel character due to the shoulder in the spectrum extending to 300 nm. See 'Materials and Methods' section for full sequences. (**b**) Sample K^+ titration. Titration is of G_3L_{444} RNA GQS with KCl additions from 0 mM to 700 mM. The arrows indicate increasing K^+ concentrations. Also included are UV thermal denaturations of G_3L_{444} RNA in 100 mM LiCl (black), NaCl (red) and KCl (blue) at (**c**) 260 nm and (**d**) 295 nm. Absorbances are normalized to the highest absorbance.

correlation is not maintained when some non-plant eukaryotes such as *M. musculus* are included, nor when regions of the *Arabidopsis* and *O. sativa* genomes are analyzed separately. The anti-correlation between GC-content of the genic and intergenic regions and their G_3 GQS density suggests possible evolutionary bias away from these potentially disrupting sequences in the coding sequence in both dicot and monocot species. It is also valid to note that guanine repeats in the genome may be associated with additional biological phenomena. For example, repeats of glycine (GGG and GGX codons) and valine, alanine, glutamate, arginine and tryptophan) (XGG or GXG codons) will have a GQS in the mRNA. Thus, GQS regions can have multiple functions.

Our study suggests that one possible function of GQS in *Arabidopsis* is regulation of large numbers of genes. The most overrepresented GO term was 'Catalytic Activity', with nearly half of the 6393 gene products annotated for catalysis encoded by genes containing at least one G_2L_{1-4} motif (Table 6). One possibility is that these GQS motifs provide a way to decrease metabolism during stress (99): in response to stressors that result in increases in cytosolic K^+ concentrations, these motifs could fold into G-quadruplexes and limit transcription and translation and thereby limit expression of enzymes. In fact, we do find that genes that are differentially expressed, either up- or downregulated when exposed to drought stress, are significantly more likely to contain a GQS sequence. This suggests GQS formation may be one of numerous mechanisms plants use to adjust to changes in environmental conditions. The most underrepresented GO terms were for tRNA, rRNA and snoRNAs, whose functions depend on proper intra- and intermolecular base pairing. The presence of GQS in these RNAs could interfere with base pairing and proper folding. One intriguing possibility is that absence of GQS could be used as a criterion in searches to identify non-coding RNAs. In addition, under different environmental conditions or in the presence of stressors, GQS formation has the potential to regulate splicing.

The G_3 GQS are enriched in the intergenic regions and depleted in the genic regions (enrichment 0.5), with the shorter loop motif, G_3L_{1-3}, depleted even further in the cds (Table 2). Increasing the number of G repeats to G_4L_{1-7} increases the potential stability of the sequence, and increases the intergenic region enrichment value from 1.8 to 2.1 while depleting the enrichment value of the genic regions from 0.5 to 0.3 (data not shown). Given

that DNA GQS reported in the literature inhibit transcription (31), the enrichments observed in the non-transcribed regions of the *Arabidopsis* genome suggest that stable GQS in these regions might be functioning to repress unwanted transcription. This idea is further supported by GQS distribution within the intergenic region, where the density of G_3L_{1-7} in non-transcribed regions is 8.7-fold higher than in intergenic transcriptional units (Table 4). The G_3L_{1-7} GQS are exceptionally stable, having melting temperatures of greater than 85 °C in physiological K^+ concentrations, supporting their ability to stay folded and thus potentially block transcription. In contrast, the G_2 GQS are enriched in the genic regions and especially the cds. These sequences have lower thermal stability and melting temperatures. One possibility is that these GQS may be more plastic, allowing switching between quadruplex and non-quadruplex structures in response to cellular conditions.

In *Arabidopsis* under unstressed conditions, the cellular K^+ concentration is around 100–150 mM (100,101). At this concentration, the most stable GQS, such as the G_3L_{221} and G_3L_{444}, which have $K^+_{1/2}$ values less than 50 mM, have the intrinsic ability to form stable quadruplex structures. Addition of flanking sequences, however, can modulate quadruplex folding if there is a competing secondary structure, as demonstrated for the G_3L_{444}+FLANK RNA oligonucleotide. Less stable GQS, such as the G_2 sequences, will most likely not be fully formed under unstressed conditions (Table 7). However, they may form during water-limiting conditions, where cellular K^+ concentrations can reach 700 mM. The potential switch in RNA structure due to cellular K^+ concentration increases could potentially affect the RNA secondary structure of a large number of genes, as up to 10 000 genes contain a G_2 GQS. The possibility for a change in RNA structure leading to a change in gene expression as a response to stress is thus present.

CONCLUSION

We have found that G-quartet sequences of varying motifs are present in *A. thaliana*. Their distribution varies with sequence motif, with G_3L_{1-7} sequences preferentially located in intergenic regions and G_2L_{1-4} sequences preferentially in genic regions. GQS located in RNA have the potential to regulate transcription and translation, perhaps as modulated by environmental and physiological conditions.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

The authors would like to thank Profs David Lilley and Paul Babitzke for helpful comments and suggestions, and Prof. Yu Zhang for helpful discussions about statistics.

FUNDING

National Science Foundation (MCB-0527102 to P.C.B.); National Science Foundation (MCB-03-45251 to S.M.A. and P.C.B.); Human Frontier Science Program (HFSP) (RGP0002/2009-C to P.C.B., S.M.A. and F.M.). Funding for open access charge: HFSP.

Conflict of interest statement. None declared.

REFERENCES

1. Smargiasso,N., Rosu,F., Hsia,W., Colson,P., Baker,E.S., Bowers,M.T., De Pauw,E. and Gabelica,V. (2008) G-quadruplex DNA assemblies: loop length, cation identity, and multimer formation. *J. Am. Chem. Soc.*, **130**, 10208–10216.
2. Morris,M.J. and Basu,S. (2009) An unusually stable G-quadruplex within the 5′-UTR of the MT3 matrix metalloproteinase mRNA represses translation in eukaryotic cells. *Biochemistry*, **48**, 5313–5319.
3. Gellert,M., Lipsett,M.N. and Davies,D.R. (1962) Helix formation by guanylic acid. *Proc. Natl Acad. Sci. USA*, **48**, 2013–2018.
4. Williamson,J.R., Raghuraman,M.K. and Cech,T.R. (1989) Monovalent cation-induced structure of telomeric DNA: the G-quartet model. *Cell*, **59**, 871–880.
5. 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.
6. Hazel,P., Huppert,J., Balasubramanian,S. and Neidle,S. (2004) Loop-length-dependent folding of G-quadruplexes. *J. Am. Chem. Soc.*, **126**, 16405–16415.
7. Paramasivan,S., Rujan,I. and Bolton,P.H. (2007) Circular dichroism of quadruplex DNAs: applications to structure, cation effects and ligand binding. *Methods*, **43**, 324–331.
8. Joachimi,A., Benz,A. and Hartig,J.S. (2009) A comparison of DNA and RNA quadruplex structures and stabilities. *Bioorg. Med. Chem.*, **17**, 6811–6815.
9. Saenger,W. (1984) *Principles of Nucleic Acid Structure*. Springer-Verlag, New York, NY.
10. Shafer,R.H. and Smirnov,I. (2000) Biological aspects of DNA/RNA quadruplexes. *Biopolymers*, **56**, 209–227.
11. Tang,C.F. and Shafer,R.H. (2006) Engineering the quadruplex fold: nucleoside conformation determines both folding topology and molecularity in guanine quadruplexes. *J. Am. Chem. Soc.*, **128**, 5966–5973.
12. Hud,N.V., Smith,F.W., Anet,F.A. and Feigon,J. (1996) The selectivity for K+ versus Na+ in DNA quadruplexes is dominated by relative free energies of hydration: a thermodynamic analysis by 1H NMR. *Biochemistry*, **35**, 15383–15390.
13. Hazel,P., Parkinson,G.N. and Neidle,S. (2006) Predictive modelling of topology and loop variations in dimeric DNA quadruplex structures. *Nucleic Acids Res.*, **34**, 2117–2127.
14. Miyoshi,D., Nakao,A. and Sugimoto,N. (2002) Molecular crowding regulates the structural switch of the DNA G-quadruplex. *Biochemistry*, **41**, 15017–15024.
15. Miyoshi,D., Matsumura,S., Nakano,S. and Sugimoto,N. (2004) Duplex dissociation of telomere DNAs induced by molecular crowding. *J. Am. Chem. Soc.*, **126**, 165–169.
16. Kumar,N. and Maiti,S. (2005) The effect of osmolytes and small molecule on quadruplex-WC duplex equilibrium: a fluorescence resonance energy transfer study. *Nucleic Acids Res.*, **33**, 6723–6732.
17. Miyoshi,D., Karimata,H. and Sugimoto,N. (2006) Hydration regulates thermodynamics of G-quadruplex formation under molecular crowding conditions. *J. Am. Chem. Soc.*, **128**, 7957–7963.
18. Kumar,N. and Maiti,S. (2008) Role of molecular crowding in perturbing quadruplex-Watson Crick duplex equilibrium. *Nucleic Acids Symp. Ser.*, 157–158.

19. Gollnick,P., Babitzke,P., Antson,A. and Yanofsky,C. (2005) Complexity in regulation of tryptophan biosynthesis in Bacillus subtilis. *Annu. Rev. Genet.*, **39**, 47–68.
20. Tucker,B.J. and Breaker,R.R. (2005) Riboswitches as versatile gene control elements. *Curr. Opin. Struct. Biol.*, **15**, 342–348.
21. Huppert,J.L. and Balasubramanian,S. (2005) Prevalence of quadruplexes in the human genome. *Nucleic Acids Res.*, **33**, 2908–2916.
22. Todd,A.K., Johnston,M. and Neidle,S. (2005) Highly prevalent putative quadruplex sequence motifs in human DNA. *Nucleic Acids Res.*, **33**, 2901–2907.
23. Azzalin,C.M., Reichenbach,P., Khoriauli,L., Giulotto,E. and Lingner,J. (2007) Telomeric repeat containing RNA and RNA surveillance factors at mammalian chromosome ends. *Science*, **318**, 798–801.
24. Huppert,J.L. and Balasubramanian,S. (2007) G-quadruplexes in promoters throughout the human genome. *Nucleic Acids Res.*, **35**, 406–413.
25. Huppert,J.L., Bugaut,A., Kumari,S. and Balasubramanian,S. (2008) G-quadruplexes: the beginning and end of UTRs. *Nucleic Acids Res.*, **36**, 6260–6268.
26. Yadav,V.K., Abraham,J.K., Mani,P., Kulshrestha,R. and Chowdhury,S. (2008) QuadBase: genome-wide database of G4 DNA–occurrence and conservation in human, chimpanzee, mouse and rat promoters and 146 microbes. *Nucleic Acids Res.*, **36**, D381–D385.
27. Rawal,P., Kummarasetti,V.B., Ravindran,J., Kumar,N., Halder,K., Sharma,R., Mukerji,M., Das,S.K. and Chowdhury,S. (2006) Genome-wide prediction of G4 DNA as regulatory motifs: role in Escherichia coli global regulation. *Genome Res.*, **16**, 644–655.
28. Thakur,R.K., Kumar,P., Halder,K., Verma,A., Kar,A., Parent,J.L., Basundra,R., Kumar,A. and Chowdhury,S. (2009) Metastases suppressor NM23-H2 interaction with G-quadruplex DNA within c-MYC promoter nuclease hypersensitive element induces c-MYC expression. *Nucleic Acids Res.*, **37**, 172–183.
29. Borgognone,M., Armas,P. and Calcaterra,N.B. (2010) Cellular nucleic-acid-binding protein, a transcriptional enhancer of c-Myc, promotes the formation of parallel G-quadruplexes. *Biochem. J.*, **428**, 491–498.
30. Cogoi,S., Paramasivam,M., Membrino,A., Yokoyama,K.K. and Xodo,L.E. (2010) The KRAS promoter responds to Myc-associated zinc finger and poly(ADP-ribose) polymerase 1 proteins, which recognize a critical quadruplex-forming GA-element. *J. Biol. Chem.*, **285**, 22003–22016.
31. 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.
32. Qin,Y., Rezler,E.M., Gokhale,V., Sun,D. and Hurley,L.H. (2007) Characterization of the G-quadruplexes in the duplex nuclease hypersensitive element of the PDGF-A promoter and modulation of PDGF-A promoter activity by TMPyP4. *Nucleic Acids Res.*, **35**, 7698–7713.
33. Cogoi,S., Paramasivam,M., Filichev,V., Geci,I., Pedersen,E.B. and Xodo,L.E. (2009) Identification of a new G-quadruplex motif in the KRAS promoter and design of pyrene-modified G4-decoys with antiproliferative activity in pancreatic cancer cells. *J. Med. Chem.*, **52**, 564–568.
34. Gonzalez,V., Guo,K., Hurley,L. and Sun,D. (2009) Identification and characterization of nucleolin as a c-myc G-quadruplex-binding protein. *J. Biol. Chem.*, **284**, 23622–23635.
35. Paramasivam,M., Membrino,A., Cogoi,S., Fukuda,H., Nakagama,H. and Xodo,L.E. (2009) Protein hnRNP A1 and its derivative Up1 unfold quadruplex DNA in the human KRAS promoter: implications for transcription. *Nucleic Acids Res.*, **37**, 2841–2853.
36. Membrino,A., Paramasivam,M., Cogoi,S., Alzeer,J., Luedtke,N.W. and Xodo,L.E. (2010) Cellular uptake and binding of guanidine-modified phthalocyanines to KRAS/HRAS G-quadruplexes. *Chem. Commun.*, **46**, 625–627.
37. Du,Z., Zhao,Y. and Li,N. (2009) Genome-wide colonization of gene regulatory elements by G4 DNA motifs. *Nucleic Acids Res.*, **37**, 6784–6798.
38. Verma,A., Yadav,V.K., Basundra,R., Kumar,A. and Chowdhury,S. (2009) Evidence of genome-wide G4 DNA-mediated gene expression in human cancer cells. *Nucleic Acids Res.*, **37**, 4194–4204.
39. Balkwill,G.D., Derecka,K., Garner,T.P., Hodgman,C., Flint,A.P.F. and Searle,M.S. (2009) Repression of translation of human estrogen receptor alpha by G-quadruplex formation. *Biochemistry*, **48**, 11487–11495.
40. 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.
41. Kumari,S., Bugaut,A. and Balasubramanian,S. (2008) Position and stability are determining factors for translation repression by an RNA G-quadruplex-forming sequence within the 5′ UTR of the NRAS proto-oncogene. *Biochemistry*, **47**, 12664–12669.
42. Arora,A., Dutkiewicz,M., Scaria,V., Hariharan,M., Maiti,S. and Kurreck,J. (2008) Inhibition of translation in living eukaryotic cells by an RNA G-quadruplex motif. *RNA*, **14**, 1290–1296.
43. 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.
44. Bock,L.C., Griffin,L.C., Latham,J.A., Vermaas,E.H. and Toole,J.J. (1992) Selection of single-stranded DNA molecules that bind and inhibit human thrombin. *Nature*, **355**, 564–566.
45. Macaya,R.F., Schultze,P., Smith,F.W., Roe,J.A. and Feigon,J. (1993) Thrombin-binding DNA aptamer forms a unimolecular quadruplex structure in solution. *Proc. Natl Acad. Sci. USA*, **90**, 3745–3749.
46. Wang,K.Y., McCurdy,S., Shea,R.G., Swaminathan,S. and Bolton,P.H. (1993) A DNA aptamer which binds to and inhibits thrombin exhibits a new structural motif for DNA. *Biochemistry*, **32**, 1899–1904.
47. 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.
48. Schaeffer,C., Bardoni,B., Mandel,J.L., Ehresmann,B., Ehresmann,C. and Moine,H. (2001) The fragile X mental retardation protein binds specifically to its mRNA via a purine quartet motif. *EMBO J.*, **20**, 4803–4813.
49. Khateb,S., Weisman-Shomer,P., Hershco-Shani,I., Ludwig,A.L. and Fry,M. (2007) The tetraplex (CGG)n destabilizing proteins hnRNP A2 and CBF-A enhance the in vivo translation of fragile X premutation mRNA. *Nucleic Acids Res.*, **35**, 5775–5788.
50. Sundquist,W.I. and Heaphy,S. (1993) Evidence for interstrand quadruplex formation in the dimerization of human immunodeficiency virus 1 genomic RNA. *Proc. Natl Acad. Sci. USA*, **90**, 3393–3397.
51. Shen,W., Gao,L., Balakrishnan,M. and Bambara,R.A. (2009) A recombination hot spot in HIV-1 contains guanosine runs that can form G-quartet structure and promote strand transfer in vitro. *J. Biol. Chem.*, **284**, 33883–33893.
52. Bugaut,A., Jantos,K., Wietor,J.L., Rodriguez,R., Sanders,J.K. and Balasubramanian,S. (2008) Exploring the differential recognition of DNA G-quadruplex targets by small molecules using dynamic combinatorial chemistry. *Angew. Chem., Int. Ed.*, **47**, 2677–2680.
53. Monchaud,D. and Teulade-Fichou,M.P. (2008) A hitchhiker's guide to G-quadruplex ligands. *Org. Biomol. Chem.*, **6**, 627–636.
54. Kikin,O., Zappala,Z., D'Antonio,L. and Bagga,P.S. (2008) GRSDB2 and GRS_UTRdb: databases of quadruplex forming G-rich sequences in pre-mRNAs and mRNAs. *Nucleic Acids Res.*, **36**, D141–D148.
55. Greenway,H. and Munns,R. (1980) Mechanisms of salt tolerance in nonhalophytes. *Annu. Rev. Plant Physiol.*, **31**, 149–190.

56. Zhang,H.X. and Blumwald,E. (2001) Transgenic salt-tolerant tomato plants accumulate salt in foliage but not in fruit. *Nat. Biotechnol.*, **19**, 765–768.
57. Swarbreck,D., Wilks,C., Lamesch,P., Berardini,T.Z., Garcia-Hernandez,M., Foerster,H., Li,D., Meyer,T., Muller,R., Ploetz,L. *et al.* (2008) The Arabidopsis Information Resource (TAIR): gene structure and function annotation. *Nucleic Acids Res.*, **36**, D1009–D1014.
58. Vogel,J.P., Garvin,D.F., Mockler,T.C., Schmutz,J., Rokhsar,D., Bevan,M.W., Barry,K., Lucas,S., Harmon-Smith,M., Lail,K. *et al.* (2010) Genome sequencing and analysis of the model grass Brachypodium distachyon. *Nature*, **463**, 763–768.
59. Adams,M.D., Celniker,S.E., Holt,R.A., Evans,C.A., Gocayne,J.D., Amanatides,P.G., Scherer,S.E., Li,P.W., Hoskins,R.A., Galle,R.F. *et al.* (2000) The genome sequence of Drosophila melanogaster. *Science*, **287**, 2185–2195.
60. Schmutz,J., Cannon,S.B., Schlueter,J., Ma,J., Mitros,T., Nelson,W., Hyten,D.L., Song,Q., Thelen,J.J., Cheng,J. *et al.* Genome sequence of the palaeopolyploid soybean. *Nature*, **463**, 178–183.
61. Sato,S., Nakamura,Y., Kaneko,T., Asamizu,E., Kato,T., Nakao,M., Sasamoto,S., Watanabe,A., Ono,A., Kawashima,K. *et al.* (2008) Genome structure of the legume, Lotus japonicus. *DNA Res.*, **15**, 227–239.
62. Young,N.D., Cannon,S.B., Sato,S., Kim,D., Cook,D.R., Town,C.D., Roe,B.A. and Tabata,S. (2005) Sequencing the genespaces of Medicago truncatula and Lotus japonicus. *Plant Physiol.*, **137**, 1174–1181.
63. Cannon,S.B., Sterck,L., Rombauts,S., Sato,S., Cheung,F., Gouzy,J., Wang,X., Mudge,J., Vasdewani,J., Schiex,T. *et al.* (2006) Legume genome evolution viewed through the Medicago truncatula and Lotus japonicus genomes. *Proc. Natl Acad. Sci. USA*, **103**, 14959–14964.
64. Wade,C.M., Kulbokas,E.J. 3rd, Kirby,A.W., Zody,M.C., Mullikin,J.C., Lander,E.S., Lindblad-Toh,K. and Daly,M.J. (2002) The mosaic structure of variation in the laboratory mouse genome. *Nature*, **420**, 574–578.
65. Opperman,C., Burke,M. and Lommel,S.A. (2007) Sequencing and analysis of the Nicotiana tabacum genome. *Recent Adv. Tob. Sci.*, **33**, 5–14.
66. Yu,J., Hu,S., Wang,J., Wong,G.K., Li,S., Liu,B., Deng,Y., Dai,L., Zhou,Y., Zhang,X. *et al.* (2002) A draft sequence of the rice genome (Oryza sativa L. ssp. indica). *Science*, **296**, 79–92.
67. Zhao,W., Wang,J., He,X., Huang,X., Jiao,Y., Dai,M., Wei,S., Fu,J., Chen,Y., Ren,X. *et al.* (2004) BGI-RIS: an integrated information resource and comparative analysis workbench for rice genomics. *Nucleic Acids Res.*, **32**, D377–D382.
68. Matsumoto,T., Wu,J.Z., Kanamori,H., Katayose,Y., Fujisawa,M., Namiki,N., Mizuno,H., Yamamoto,K., Antonio,B.A., Baba,T. *et al.* (2005) The map-based sequence of the rice genome. *Nature*, **436**, 793–800.
69. Ouyang,S., Zhu,W., Hamilton,J., Lin,H., Campbell,M., Childs,K., Thibaud-Nissen,F., Malek,R.L., Lee,Y., Zheng,L. *et al.* (2007) The TIGR Rice Genome Annotation Resource: improvements and new features. *Nucleic Acids Res.*, **35**, D883–D887.
70. Tuskan,G.A., Difazio,S., Jansson,S., Bohlmann,J., Grigoriev,I., Hellsten,U., Putnam,N., Ralph,S., Rombauts,S., Salamov,A. *et al.* (2006) The genome of black cottonwood, Populus trichocarpa (Torr. & Gray). *Science*, **313**, 1596–1604.
71. Rensing,S.A., Lang,D., Zimmer,A.D., Terry,A., Salamov,A., Shapiro,H., Nishiyama,T., Perroud,P.F., Lindquist,E.A., Kamisugi,Y. *et al.* (2008) The Physcomitrella genome reveals evolutionary insights into the conquest of land by plants. *Science*, **319**, 64–69.
72. Paterson,A.H., Bowers,J.E., Bruggmann,R., Dubchak,I., Grimwood,J., Gundlach,H., Haberer,G., Hellsten,U., Mitros,T., Poliakov,A. *et al.* (2009) The Sorghum bicolor genome and the diversification of grasses. *Nature*, **457**, 551–556.
73. Jaillon,O., Aury,J.M., Noel,B., Policriti,A., Clepet,C., Casagrande,A., Choisne,N., Aubourg,S., Vitulo,N., Jubin,C. *et al.* (2007) The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature*, **449**, 463–467.
74. Schnable,P.S., Ware,D., Fulton,R.S., Stein,J.C., Wei,F., Pasternak,S., Liang,C., Zhang,J., Fulton,L., Graves,T.A. *et al.* (2009) The B73 maize genome: complexity, diversity, and dynamics. *Science*, **326**, 1112–1115.
75. Rajendran,A., Nakano,S. and Sugimoto,N. (2010) Molecular crowding of the cosolutes induces an intramolecular i-motif structure of triplet repeat DNA oligomers at neutral pH. *Chem. Commun.*, **46**, 1299–1301.
76. Maere,S., Heymans,K. and Kuiper,M. (2005) BiNGO: a Cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks. *Bioinformatics*, **21**, 3448–3449.
77. Shannon,P., Markiel,A., Ozier,O., Baliga,N.S., Wang,J.T., Ramage,D., Amin,N., Schwikowski,B. and Ideker,T. (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.*, **13**, 2498–2504.
78. Cline,M.S., Smoot,M., Cerami,E., Kuchinsky,A., Landys,N., Workman,C., Christmas,R., Avila-Campilo,I., Creech,M., Gross,B. *et al.* (2007) Integration of biological networks and gene expression data using Cytoscape. *Nat. Protoc.*, **2**, 2366–2382.
79. Ashburner,M., Ball,C.A., Blake,J.A., Botstein,D., Butler,H., Cherry,J.M., Davis,A.P., Dolinski,K., Dwight,S.S., Eppig,J.T. *et al.* (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.*, **25**, 25–29.
80. Sosnick,T.R. (2001) Characterization of tertiary folding of RNA by circular dichroism and urea. *Curr. Protoc. Nucleic Acid Chem.*, **4**, 11.5.1–11.5.10.
81. Mergny,J.L., Phan,A.T. and Lacroix,L. (1998) Following G-quartet formation by UV-spectroscopy. *FEBS letters*, **435**, 74–78.
82. Wieland,M. and Hartig,J.S. (2007) RNA quadruplex-based modulation of gene expression. *Chem. Biol.*, **14**, 757–763.
83. Simoens,C.R., Gielen,J., Van Montagu,M. and Inze,D. (1988) Characterization of highly repetitive sequences of Arabidopsis thaliana. *Nucleic Acids Res.*, **16**, 6753–6766.
84. Richards,E.J., Goodman,H.M. and Ausubel,F.M. (1991) The centromere region of Arabidopsis thaliana chromosome 1 contains telomere-similar sequences. *Nucleic Acids Res.*, **19**, 3351–3357.
85. Lee,C., Sasi,R. and Lin,C.C. (1993) Interstitial localization of telomeric DNA sequences in the Indian muntjac chromosomes: further evidence for tandem chromosome fusions in the karyotypic evolution of the Asian muntjacs. *Cytogenet. Cell Genet.*, **63**, 156–159.
86. Matsui,A., Ishida,J., Morosawa,T., Mochizuki,Y., Kaminuma,E., Endo,T.A., Okamoto,M., Nambara,E., Nakajima,M., Kawashima,M. *et al.* (2008) Arabidopsis transcriptome analysis under drought, cold, high-salinity and ABA treatment conditions using a tiling array. *Plant Cell Physiol.*, **49**, 1135–1149.
87. Yamada,K., Lim,J., Dale,J.M., Chen,H., Shinn,P., Palm,C.J., Southwick,A.M., Wu,H.C., Kim,C., Nguyen,M. *et al.* (2003) Empirical analysis of transcriptional activity in the Arabidopsis genome. *Science*, **302**, 842–846.
88. Halder,K., Wieland,M. and Hartig,J.S. (2009) Predictable suppression of gene expression by 5′-UTR-based RNA quadruplexes. *Nucleic Acids Res.*, **37**, 6811–6817.
89. English,A.C., Patel,K.S. and Loraine,A.E. (2010) Prevalence of alternative splicing choices in Arabidopsis thaliana. *BMC Plant Biol.*, **10**, 102.
90. Lazar,G. and Goodman,H.M. (2000) The Arabidopsis splicing factor SR1 is regulated by alternative splicing. *Plant Mol. Biol.*, **42**, 571–581.
91. Palusa,S.G., Ali,G.S. and Reddy,A.S. (2007) Alternative splicing of pre-mRNAs of Arabidopsis serine/arginine-rich proteins: regulation by hormones and stresses. *Plant J.* **49**, 1091–1107.
92. Eckardt,N.A. (2002) Alternative splicing and the control of flowering time. *Plant Cell*, **14**, 743–747.

93. Reddy,A.S. (2007) Alternative splicing of pre-messenger RNAs in plants in the genomic era. *Annu. Rev. Plant Biol.*, **58**, 267–294.
94. Cantor,C.R. and Schimmel,P.R. (1980) *Biophysical Chemistry, part II: Techniques for the Study of Biological Structure and Function*. W.H. Freeman, San Francisco, CA.
95. Zuker,M. (2003) Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res.*, **31**, 3406–3415.
96. Deng,H. and Braunlin,W.H. (1995) Duplex to quadruplex equilibrium of the self-complementary oligonucleotide d(GGGGCCCC). *Biopolymers*, **35**, 677–681.
97. Pal,M., Ponticelli,A.S. and Luse,D.S. (2005) The role of the transcription bubble and TFIIB in promoter clearance by RNA polymerase II. *Mol. Cell*, **19**, 101–110.
98. Yonaha,M. and Proudfoot,N.J. (1999) Specific transcriptional pausing activates polyadenylation in a coupled in vitro system. *Mol. Cell*, **3**, 593–600.
99. Achard,P., Cheng,H., De Grauwe,L., Decat,J., Schoutteten,H., Moritz,T., Van Der Straeten,D., Peng,J. and Harberd,N.P. (2006) Integration of plant responses to environmentally activated phytohormonal signals. *Science*, **311**, 91–94.
100. Leigh,R.A. and Wyn Jones,R.G. (1984) A hypothesis relating critical potassium concentrations for growth to the distribution and functions of this ion in the plant cell. *New Phytol.*, **97**, 1–13.
101. Leigh,R.A. (2001) Potassium homeostasis and membrane transport. *J. Plant Nutr. Soil Sci.*, **164**, 193–198.