Turkish Journal of Botany

Volume 43 | Number 2 Article 2

1-1-2019

G-Quadruplex enrichment analysis reveals their role as intronic regulatory elements in plants

OSMAN DOLUCA

TÜBİTAK

Turkish Journal of Botany

http://journals.tubitak.gov.tr/botany/

Research Article

Turk J Bot
(2019) 43: 151-166
© TÜBİTAK
doi:10.3906/bot-1808-25

G-Quadruplex enrichment analysis reveals their role as intronic regulatory elements in plants

Osman DOLUCA*
Department of Biomedical Engineering, Faculty of Engineering, İzmir University of Economics, İzmir, Turkey

Received: 07.08.2018 • **Accepted/Published Online:** 21.01.2019 • **Final Version:** 03.07.2019

Abstract: G-Quadruplexes, a class of noncanonical but highly stable nucleic acid structures, have the potential to be part of the regulatory mechanism of cells. They can form in the genome where the double-stranded helix is unwound to facilitate formation of a G-quadruplex. The biological significance of these structures is yet to be understood entirely. This work presents a novel approach and investigates common characteristics in the distribution of G-quadruplexes relative to genes in plants through analysis of genomes and gene expressions. The results indicate that G-quadruplex distribution has gone through significant changes with the evolution of higher plants and, for the first time, that G-quadruplexes enriched at the beginning of introns may have a regulatory role during transcription.

Key words: G-Quadruplex, plant genetics, intron-mediated enhancement, epigenetics

1. Introduction

G-Quadruplexes (G4s) are a class of noncanonical and four-stranded DNA structures with roles revealed especially in chromatin maintenance and gene regulation (Burge et al., 2006). The first observations of G4s in biological samples were at telomeres where repeating G4s were demonstrated to prevent telomere elongation (Fletcher et al., 1998). It is not only telomeres where G4s were observed. Many nontelomeric G4s were found, surprisingly and more frequently, in genes associated with cancer and regulation (Eddy and Maizels, 2006; Huppert and Balasubramanian, 2007). The first indication of the role of G4s in gene regulation was observed in the c-myc promoter, where G-quadruplex disruptive mutations in the promoter lead to overexpression of c-myc (Grand et al., 2004). This was followed by several other promoter G4s (Cogoi and Xodo, 2006; Fernando et al., 2006; Sun et al., 2008), yet the regulatory effects of the G4s are not only limited to the promoter and associated with transcription initiation. They are also observed in UTRs and are associated with translation (Huppert et al., 2008; Beaudoin and Perreault, 2013; Kwok et al., 2015).

Facing such a large variety of modes of regulation, systematic analysis of genomes is required to distinguish G4s and their associated modes of regulation. Several genomewide analyses were performed to reveal G4 enrichment in the regulatory regions in humans (Huppert, 2005; Todd et al., 2005), several animals (Zhao et al., 2007; Verma et al., 2008), fungi (Hershman et al., 2008), prokaryotes (Wieland and Hartig, 2009; Kaplan et al., 2016), and viruses (Biswas et al., 2016). In the case of plants, genome-wide analysis was performed for only five species, *Arabidopsis thaliana* (Mullen et al., 2010; Takahashi et al., 2012), *Oryza sativa* (Takahashi et al., 2012; Wang et al., 2015), *Populus trichocarpa* (Takahashi et al., 2012), *Vitis vinifera* (Takahashi et al., 2012), and *Zea mays* (Andorf et al., 2014), while many more plant species have been sequenced and annotated and are available online for kingdom-wide G4 analysis (Andorf et al., 2014; Proost et al., 2014). Moreover, all genome-wide analysis studies except one focused on only two gene features, the transcription start site (TSS) and transcription termination site (TTS), and the relative positions of G4s. Recently, a larger study was performed among plant genomes, focusing on gene ontology of G4-related genes (Garg et al., 2016). G4s were studied according to their genomic region, however, and not their relative positions in those regions.

Here, 23 plants including mosses, algae, and higher plants have been analyzed for G-quadruplex distribution relative to gene features: TSS, TTS, translation start codon (AUG), translation stop codon (STOP), first exon–intron boundary (EXINT), and first intron–exon boundary (INTEX). These distributions are being reported for the first time and for a large number of plant species.

A new approach is also proposed to determine G4 distribution. Instead of considering the center of a G4,

* Correspondence: osman.doluca@ieu.edu.tr

every nucleotide that participates in the G4, including the loops, are considered for the distribution. This approach provides a fair representation of lengthy G4s, since longer G4-forming sequences have greater impact on DNA topology, especially in terms of torsion. This approach also helps form distinguished enrichment peaks when G4s overlap for a portion of their length, even if their central relative position is different. It should be noted that such an approach does not represent the guanine density and thus G4 density; however, the aim of this approach is to study the enriched positions of G4s, not the G4 density.

These distributions revealed G4-enriched hotspots conserved in the vicinity of gene features. These hotspots are further analyzed with a novel approach. The coexpression of *O. sativa* genes that contain a G4 at a particular hotspot is analyzed and clustered for the first time and the results provide the first indication of the role of G-quadruplexes in intron-mediated enhancement in plants.

2. Materials and methods

2.1. G-Quadruplex prediction

Genome sequences of 23 different plant species were downloaded in FASTA format from the Plaza 3.0 website (Proost et al., 2014). The list of the plant species and their abbreviations used in this paper may be found in Table 1. The genomes were previously assembled from v.1.0 assembly for aly (Hu et al., 2011), TAIR10 assembly for ath (The Arabidopsis Genome Initiative, 2000), RefBeet 1.1 for bvu (Dohm et al., 2014), Melonomics v3.5 for cme (Garcia-Mas et al., 2012), JGI v5.0 assembly and annotation v5.3.1 based on Augustus u11.6 for cre (Merchant et al., 2007), JGI annotation v1.0 on assembly v1 for cru (Slotte et al., 2013), JGI v1 assembly and v1.0 annotation for csi (Xu et al., 2013), JGI annotation v1.1 on assembly v1.0 for egr (Myburg et al., 2014), JGI Glyma1.1 annotation of the chromosome-based Glyma1 assembly for gma (Schmutz et al., 2010), JGI annotation v2.1 on assembly v2.0 for gra (Wang et al., 2012), Cassava4 for mes (Prochnik et al., 2012), JCVI 4.0 for mtr (Young et al., 2011), JGI v2.0 assembly and annotation for olu (Palenik et al., 2007), MSU RGAP 7 for osa (International Rice Genome Sequencing Project, 2005), JGI assembly release v1.1 and COSMOSS annotation v1.6 for ppa (International Rice Genome Sequencing Project, 2005; Rensing et al., 2008), JGI release v1.0 for ppe (Verde et al., 2013), JGI assembly release v3.0 and annotation v3.0 for ptr (Tuskan et al., 2006), JCVI 1.0 for rco (Chan et al., 2010), ITAG 2.3 for sly (Tomato Genome Consortium, 2012), D. Gilbert public gene set 8 Mar 2012 on assembly v1.1 for tca (Argout et al., 2011), genescope v1 for vvi (Jaillon et al., 2007), and assembly B73 for zma (Schnable et al., 2009).

Each whole genome sequence was initially scanned for putative G-quadruplex-forming sequences using the quadparser algorithm (Huppert, 2005). The G4 prediction results included the sequences, the chromosome numbers, the coordinates, and the strands of the predicted G4s. The algorithm discovers any sequence matching the $G \geq 3$ $(N1 - 7G \geq 3) \geq 3$ pattern and concurrently lists 4 or more consecutive G-tracts of a minimum length of 3 guanines with a minimum distance of 1 and maximum distance of 7 bases as a putative G-quadruplex. These criteria count nonoverlapping predictions as a single prediction as long as the distance between the closest G-tracts is less than 7 bases. This is particularly useful since it is not known which G-tracts would assemble to form G4s.

2.2. G4-Participating nucleotide distribution analysis

Structural annotations of the plant genomes were downloaded from the Plaza 3.0 website (Proost et al., 2014) in comma-separated value (csv) format. Positions of four features, transcription start site (TSS), first base of the start codon (AUG), first base of the stop codon (STOP), and transcription termination site (TTS), were located for each protein-coding gene if the untranslated regions were clearly annotated. Any genes without UTR regions were discarded for this analysis. Among selected genes, positions of the first splicing donor site (EXINT) and first splicing acceptor site (INTEX) features were also listed. Genes without annotated introns were discarded for EXINT and INTEX analysis. During the listing of the features, the directionality and the strand of the gene were considered.

Distribution analysis evaluated each nucleotide of each putative G-quadruplex, regardless of the type of the nucleotide, and checked whether it was present within 1000 bases of a feature of focus of any gene. The hit was then recorded on a G4-participating nucleotide distribution profile spectrum spanning from –1000 to +1000 in relation to the directionality of the gene with the origin corresponding to the feature. Any hit more than 1000 bases from a feature was dismissed. Whether the strand of the hit was the same as the template strand or the coding strand of the gene, the hit was recorded on the template strand profile or the coding strand profile, respectively. This resulted in the creation of two separate profiles per feature.

2.3. Formation of expression similarity and P-value matrices

The identities of the genes associated with the G4s found at each hotspot or negative control was used to extract the expression dataset from the next-generation sequencing transcriptome data of the Rice Genome Annotation Project database (Kawahara et al., 2013), in which sequence reads from the NCBI Sequence Read Archive were mapped and expression values were calculated as described. All nonquantitative and categorical data were omitted to simplify the computations, leaving expression abundance data from 16 different experiments (He et al., 2010; Zemach et al., 2010; Davidson et al., 2012). Any

Table 1. List of abbreviations, names, number of predicted G-quadruplexes, G-quadruplex densities, number of genes with annotated UTRs, and number of genes with defined introns for the species analyzed in this study.

Abbreviation	Species	Vernacular name*	No. of putative G4s found	G4 density (G4 per million bp)**	No. of genes used with defined UTRs	No. of genes used for intron/exon boundaries
aly	*Arabidopsis lyrata*	Lyrate cress	2142	12	20287	14372
ath	*Arabidopsis thaliana*	Thale cress	1232	10	18847	15139
bvu	*Beta vulgaris*	Sugar beet	12555	52	23279	17086
cme	*Cucumis melo*	Muskmelon	13936	21	11891	9602
cre	*Chlamydomonas reinhardtii*	Green algae	105325	982	17693	16348
cru	*Capsella rubella*	Pink shepherd's purse	2261	17	11869	9946
csi	*Citrus sinensis*	Sweet orange	5864	23	10456	8914
egr	*Eucalyptus grandis*	Rose gum	30884	48	18258	13593
gma	*Glycine max*	Soybean	20191	21	39899	31604
gra	*Gossypium raimondii*	Cotton	21241	28	26930	21985
mes	*Manihot esculenta*	Cassava	6483	15	9533	7908
mtr	*Medicago truncatula*	Barrel medic	8106	21	18841	14549
olu	*Ostreococcus lucimarinus*	Green algae	953	72	761	189
osa	*Oryza sativa ssp. japonica*	Rice	40779	109	20929	15727
ppa	*Physcomitrella patens*	Moss	12137	27	11049	8800
ppe	*Prunus persica*	Peach	9025	40	6368	5230
ptr	*Populus trichocarpa*	Black cottonwood	9027	21	24589	19361
rco	*Ricinus communis*	Castor bean	4034	12	2524	1961
sly	*Solanum lycopersicum*	Tomato	20059	27	10577	10089
tca	*Theobroma cacao*	Cacao	5654	17	28836	21977
tpa	*Thellungiella parvula*	Saltwater cress	711	6	11525	11524
vvi	*Vitis vinifera*	Grape vine	17472	37	10811	10721
zma	*Zea mays*	Maize	150001	73	26726	18560

*Vernacular names were obtained from NCBI, Taxonomy browser, or Plaza 3.0 as possible.
**G4 density is calculated based on the number of putative G4s found and the total number of nonambiguous nucleotides (A, T, C, and G, excluding N) since ambiguous nucleotides cannot be analyzed for G4 prediction.

gene listed with no recorded expression for any of the experiments was also omitted from the expression dataset.

For each hit in a selected hotspot or negative control, the names of the genes and the G4 sequences were listed. The gene names were then used to extract expression abundances from the next-generation sequencing transcriptome data and G4-associated gene expression data were then used to construct an expression similarity matrix using Pearson correlation between each gene pair, scoring between –1 and 1 using Eq. (1), where a and b are expression vectors of genes and are the means of their entries. Correlation was then converted to distance using Eq. (2). The distances were used for hierarchical clustering using UPGMA linkage and formed a dendrogram for each hotspot (Michener and Sokal, 1958).

$$r(a,b) = \frac{(a-\bar{a})\cdot(b-\bar{b})}{\|a-\bar{a}\|_2\|b-\bar{b}\|_2} \quad (1) \qquad d(a,b) = 1 - r(a,b) \quad (2)$$

Concurrently, a P-value matrix was calculated from the correlations, indicating the probability of random occurrence of the associated correlation. For each correlation, a 2-tailed P-value was calculated and placed in the corresponding location in a P-value matrix. The number of gene pairs with a correlation coefficient higher than 0.7 and a P-value lower than 0.01 is reported as correlated gene pair count.

The expression similarity matrix and P-value matrix were then reorganized to align with the corresponding leaves of the dendrogram for visualization purposes.

Pearson correlations and P-values were calculated in the Python environment using the scipy module (version 0.19.1). Dendrograms were visualized using the matplotlib package (version 2.0.2). For each hotspot the number of gene pairs that showed expression similarity value above 0.7 and P-value below 0.01 were counted. This value is reported as the number of correlated gene pairs.

To be able to evaluate the significance of these counts and to eliminate doubt that any randomly selected gene set may also have a similar level of coregulation, 100 sets of randomly picked genes were generated from the identical expression dataset to simulate the population of all possible expression similarity matrices with the same size. Each random set contained an equal number of randomly selected genes to the number of G4-associated genes used for particular hotspot testing. For each random set, expression similarity and P-value matrices were constructed. As applied to the former set, the numbers of gene pairs with a correlation coefficient of at least 0.7 and a maximum P-value of 0.01 per random set were calculated. This process was repeated for each hotspot or negative control since each had a different number of genes. From these data a z-score for each hotspot is reported, indicating significances of the correlated gene pair count (Table 2).

2.4. G-Quadruplex identity matrix construction

For each hotspot or negative control, the nucleotide hit corresponded to a particular G-quadruplex. These G-quadruplexes may share similarity in topology, which requires biophysical experimentation. Alternatively, the similarities of these G-quadruplexes may be captured through pairwise comparison of the sequences that form the G-quadruplex as we have suggested previously (Kaplan et al., 2016). Each G-quadruplex-forming sequence in a hotspot or a negative control was compared using pairwise alignment using the biopython package with local alignment function and a match score of 1 and mismatch, gap opening, and gap extension penalties of –2, –4, and –0.5, respectively. The highest score from each alignment was then divided by the length of the shorter sequence since that corresponds to the highest score if the sequence were to be aligned with itself. The resulting value was equal to 1 or lower, 1 indicating an exact match of the shorter sequence with a portion of the longer sequence. This calculation method ensured that even if the two sequences were not equal in size, the shorter one could have a sequence identical to a portion of the larger one and thus both could form identical G-quadruplexes. These scores were placed in a G-quadruplex identity matrix corresponding to the expression similarity matrix.

All analyses were performed in Python 2.7.12 using precompiled numpy+mkl (version 1.13.1) and scipy (version 0.19.1) packages. Other dependencies listed above were installed through pip and used up to date.

3. Results

3.1. Transcription start site-aligned (TSS-aligned) distribution profiles

All species showed G-quadruplex enrichment at or close to the TSS feature on the template strand (Figure 1). Only moss and algae species, ppa, olu, and cre, seemed to have significant G4 formation on the coding strand in comparison to their own template strands. It is important to note that these three species are considered lower plants (moss and algae) and are distant on the evolutionary tree from the others analyzed (Figure 2) (Eddy et al., 1992, Sayers et al. 2009). However, the profiles of the coding strands of these species showed difference between each other.

Within *Arabidopsis* species, aly and ath, a single peak upstream of the feature was observed while within Fabaceae members, mtr and gma, increased G4 formation centered around 250 bp downstream was detected, distinct from the other species. Within *Malpighiales*, mes, ptr, and rco, the similarity was unclear due to the low number of G4 hits in the rco profile. This was probably due to the low number of available genes used in its analysis instead of the number of G4 predictions (Table 1).

Table 2. G-Quadruplex-enriched hotspots and negative controls for *O. sativa* (osa) genome chosen for analysis. First four hotspots are named according to their aligned feature. Last two negative controls are chosen due to absence of any significant G4 enrichment. The positions and strands relative to the given feature are given in the second column. Numbers of correlated gene pairs are given in bold if statistically significant.

Hotspot or negative control	Hotspot or negative control and strand	G4-associated gene count	Correlated gene pairs	z-score	P-value*
HS1-TSS	–52 template	262	3677 (P < 0.05)	1.70	0.04557
HS2-TSS	–9 template	350	8827 **(P < 0.01)**	4.60	**<0.00001**
HS3-AUG	–16 coding	147	1411 **(P < 0.01)**	2.80	**0.00256**
HS4-EXINT	+24 template	332	10139 **(P < 0.01)**	11.59	**<0.00001**
CS5-CTRL	–1000 coding	26	50 (P > 0.01)	1.60	0.05480
CS6-CTRL	+1000 template	31	22 (P > 0.01)	–1.49	0.93189
*Based on right-tailed z-test.					

On the other hand, higher plant species tended to show a common characteristic in the template strand: an increase in G4 formation towards the feature at the center. Although the position and the number of local maxima changed from plant to plant, they were accumulated in the vicinity of the TSS feature. This peak was broad and several hundred base pairs in width for zma, cme, osa, gra, ptr, tca, egr, vvi, ppe, and gra consisting of several local maxima. For Fabaceae, gma and mtr, as well as for tpa, the broad peak was shifted further downstream. *Arabidopsis* members aly and ath along with sly showed a low-angle peak at about 65 bp upstream of the TSS feature with a baseline width of 100 bp.

Local maxima at the three coordinates seemed to be conserved in most template strand profiles of higher plants: 60 bp (±7) upstream, –10 bp (±12) upstream, and a broad 50 bp (±15) downstream. The first was seen in zma, cme, osa, ptr, sly, aly, ath, and vvi. The second sharp spike was observed in ptr, tca, mtr, gra, gma, osa, egr, ppa, ptr, and cme. The third peak tended to be broader and clear in zma, ptr, osa, tca, and cme.

Poaceae, zma and osa, showed remarkable similarity on the coding strand profiles with the broad peak on the downstream flank and the smaller upstream. A peak right at the TSS feature was distinguished for gma, gra, and egr.

3.2. Distribution profiles for first codon (AUG)

The majority of AUG-aligned distribution profiles showed a common feature on the template strand: a broad peak split into two with a sharp dip at the AUG feature. This split broad peak characteristic is an indication that the start codon cannot tolerate G4 formation. This finding was supported by a previous study on maize (zma) (Andorf et al., 2014). Among lower plants, cre and ppa also showed the dip at the feature, supporting intolerance of G4 formation at the AUG feature in the template strand. In the case of cre, the dip was also evident with gradual decrease from the flanks rather than a sharp dip (Figure 3).

It should be mentioned that, as in TSS-aligned distribution profiles, the coding strand showed lower G4 formation in AUG-aligned profiles. On the coding strand, a similar dip at the feature exists; however, it is more obscure. This is expected as G4 formation would block the recognition of the AUG feature. While some species showed increase in G4 formation before the feature within the vicinity of the start codon on the coding strand (vvi, sly, osa, ppa, egr, ath, cru, cme), others showed accumulated G4 formation after the dip (aly, cre, mtr, ptr, ppe, rco). Some species showed local maxima on both sides of the feature (gma, gra, olu, tpa, tca, zma). The upstream peak was especially intriguing in osa since this peak was the only maximum lying upstream of a plateau that started from the AUG.

3.3. Distribution profiles for first splicing donor site (EXINT)

In the EXINT-aligned G4-participating nucleotide distribution profiles for both template and coding strands, there were increased numbers of G4s upstream of the feature forming a broad peak, except in *Chlorophyta*, olu and cre. At the same time, prominently, most of the template profiles showed a sharp peak between 5 and 80 bases downstream with a maximum at around 23 (±3) bases downstream. This peak clearly points out the boundary of the first exon and the first intron as the slope was sharp and the distance between the beginning of the peak and the feature was consistent. It should also be noted that cre showed a distinct downstream peak on the coding strand profile at 28 bases downstream (Figure 4).

3.4. Distribution profiles for first splicing acceptor site (INTEX)

In general, the transition from the first intron to the second exon showed a decrease in G4 formation, with minima at the INTEX feature both on coding and template strands for most profiles. *Poaceae*, osa and zma, showed a distinct

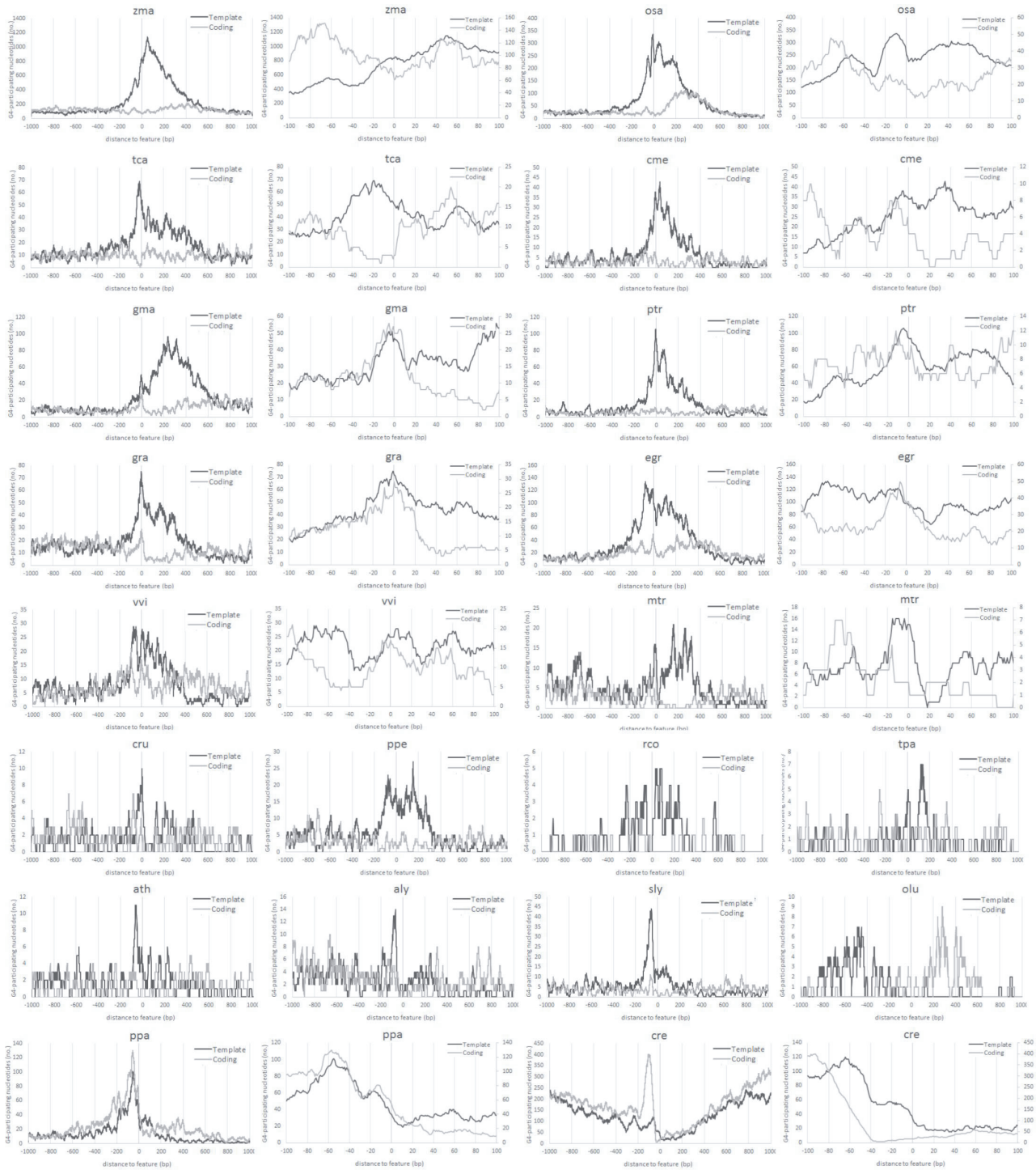

Figure 1. G4-Participating nucleotide distribution profiles in the ±1000 (left) or ±100 (right) nucleotides vicinity of TSS separately for coding (light gray) and template (dark gray) strands.

peak at around –69 preceding a sharp drop in the template strand. Different from the rest of the profiles, the template strand profile of cre had a distinct peak right before the feature with a maximum at –25 (Figure 5).

3.5. Distribution profiles for stop codon (STOP)

G4 enrichment was not very pronounced for stop codon-aligned distribution profiles; however, all profiles showed a common dip at the STOP feature. It should be noted that the

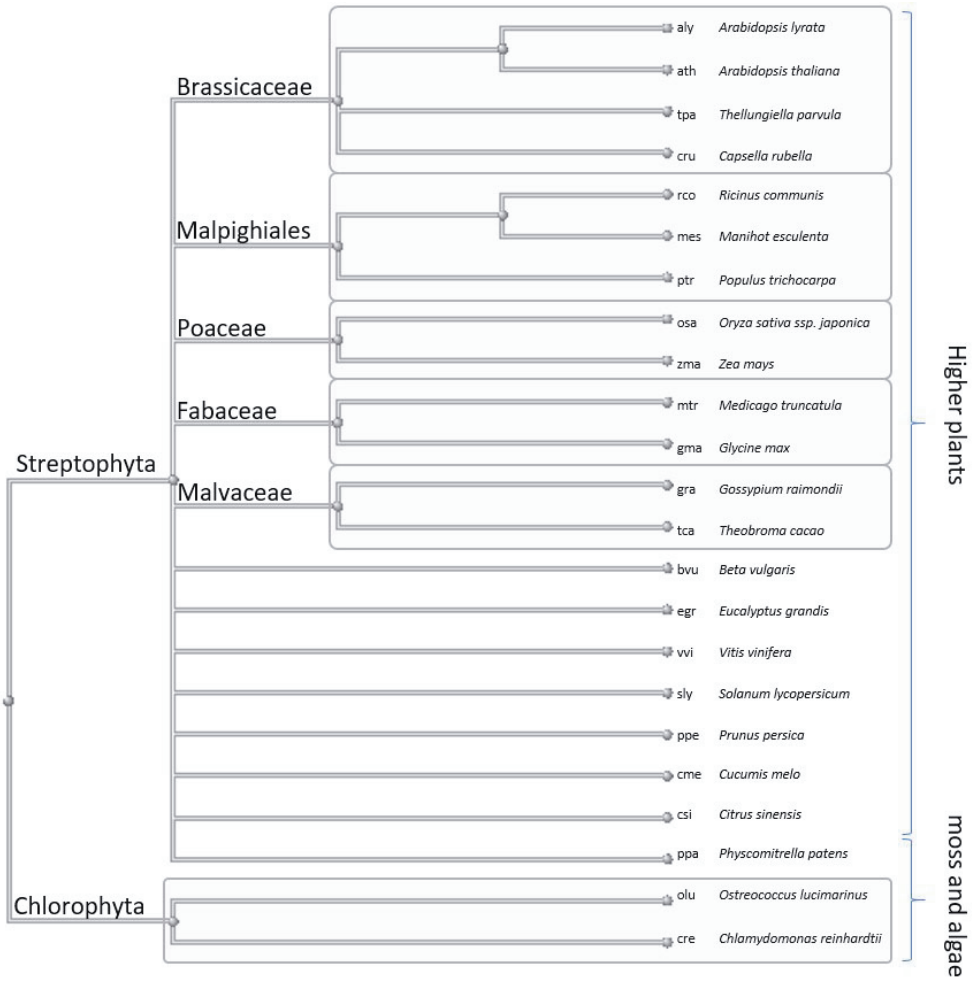

Figure 2. Evolutionary tree of plants according to NCBI taxonomy database.

number of genes with G4s in the proximity of this feature was much less than in other profiles, with the exception of cre, as was often the case in its profiles. For coding strands and inherently for the transcripts, a sharp increase in G4 formation starting from STOP clearly separated the open reading frame from the 3′UTR for most plants. Since this characteristic results in G4 formation at the mRNA level, it could also be related to the regulation of the translational termination. This is even more noticeable in cre (Figure 6).

3.6. Distribution profiles for transcription termination site (TTS)

Transcription termination site-aligned distribution profiles showed a common gradual decrease towards the center in the template strand of *Poaceae*, zma and osa. On the other hand, most profiles showed a sharp increase at the feature (gra, gma, egr, ptr, sly, cre). On the coding strand, similar profiles were observed with increased G4s either at the center or downstream. These characteristics were pronounced for egr, gra, gma ptr, and cre only (Figure 7).

3.7. Investigation of regulatory roles of enriched G4s

Two properties of regulatory elements were expected to be conserved: their topology (in duplex or quadruplex form) and location. In terms of topology, although the types of G-quadruplexes formed are not clear to us at this point, it was already predicted that these elements are forming some type of four-stranded kink, or G-quadruplex. In terms of location, specific positions among the profiles across species were already observed to be abundant in G4s and are listed in Table 1. Since G4s at close coordinates showed similarity in both topology and location, analysis of G4s located at hotspots for coexpression of their associated genes would enlighten us about the strength of influence of G4s on coregulation as well as the mechanism by which the coregulation occurs.

For that purpose, four specific locations were selected due to enriched G4 formation and referred to as hotspots. Further analysis was performed to comprehend the level of regulatory impact of the G4s at these hotspots. For

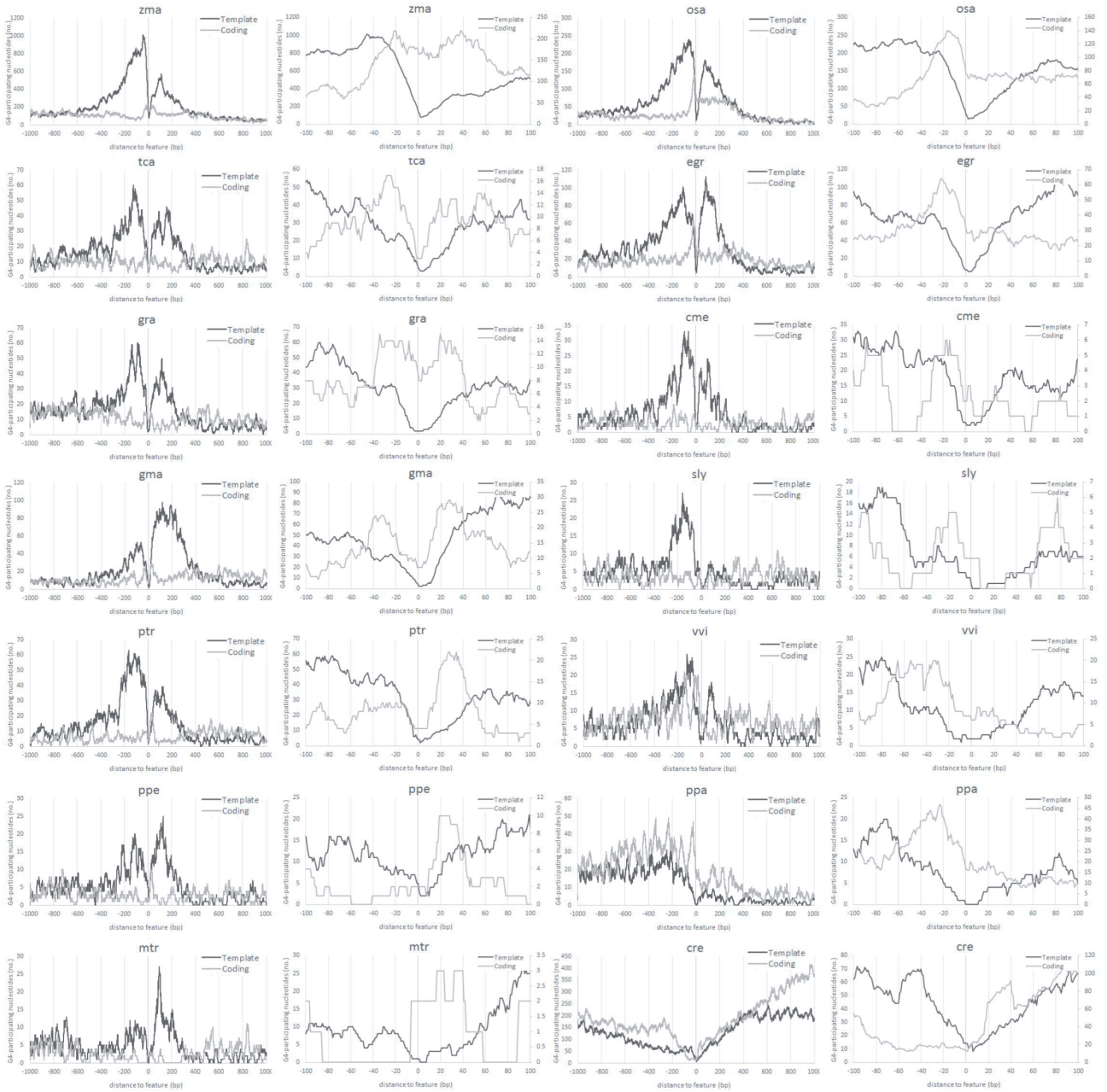

Figure 3. G4-Participating nucleotide distribution profiles in the ±1000 (left) or ±100 (right) nucleotides vicinity of AUG separately for coding (light gray) and template (dark gray) strands.

this analysis the coexpression data were best presented in the literature by the Rice Genome Annotation Project in the form of an assembled coexpression table for a large number of gene loci derived from the NCBI Sequence Read Archive (Kawahara et al., 2013). Fortunately, rice (osa) has shown several conserved G4 enrichment spots in the distribution profiles.

The four G4 hotspots on osa chosen for the analysis were –52 of the TSS at the template strand (HS1-TSS), –9 of the TSS at the template strand (HS2-TSS), –16 of the AUG at the coding strand (HS3-AUG), and 24 of EXINT at the template strand (HS4-EXINT) (Figure 8; Table 2). All of these hotspots were chosen for their unmistakable presence and conservation across species.

HS1-TSS, a G4 hotspot located at –52 of osa TSS, had higher than average number of correlated gene pairs for any random gene set with the same number of genes. However, it showed lower statistical significance ($P < 0.05$)

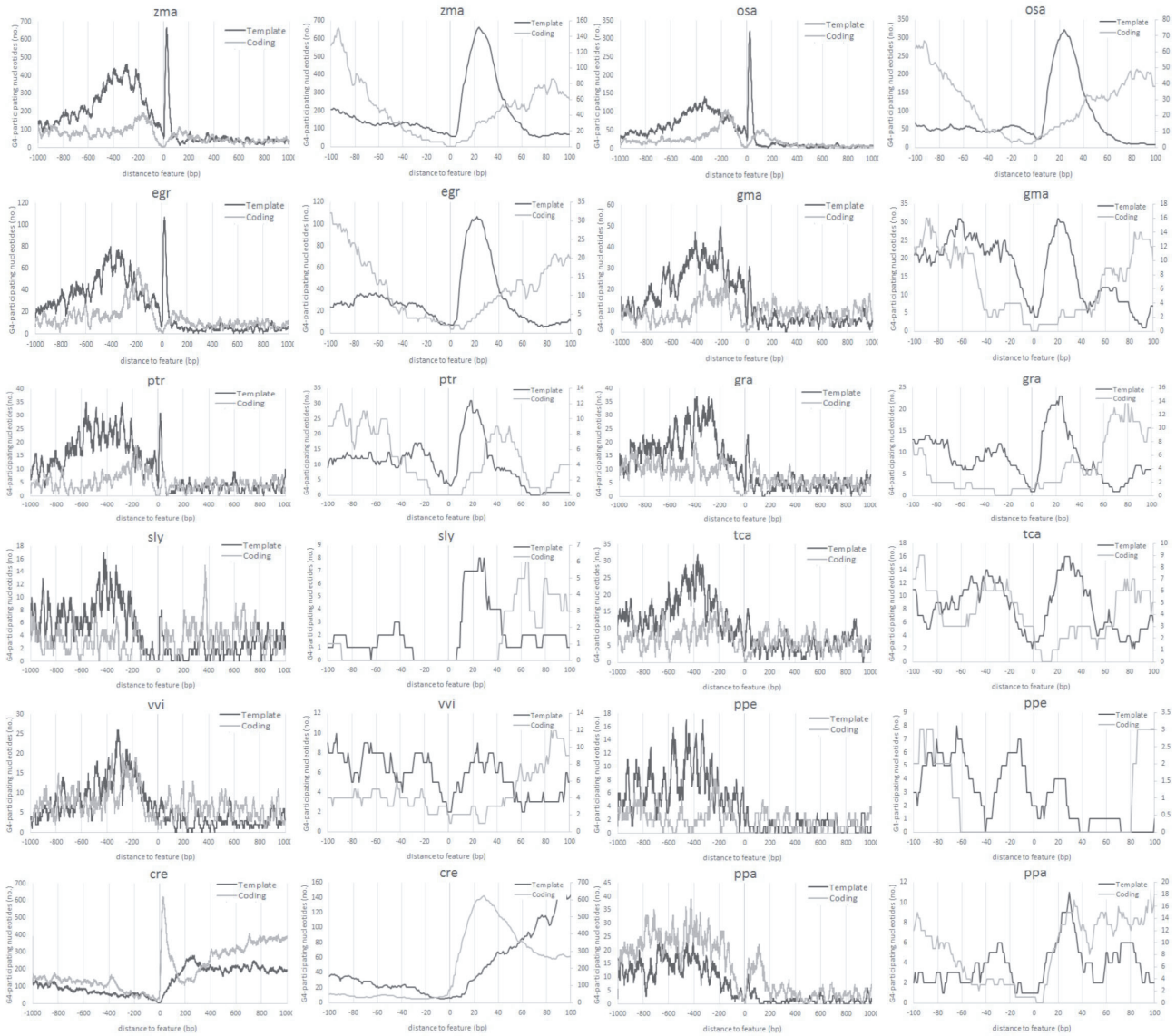

Figure 4. G4-Participating nucleotide distribution profiles in the ±1000 (left) or ±100 (right) nucleotides vicinity of EXINT separately for coding (light gray) and template (dark gray) strands.

than other G4-enriched hotspots. It is important to note that this peak was not as pronounced as others since it was part of a broader peak with a maximum at –9 where another hotspot was chosen. HS2-TSS at –9 of the template strand had a very high number of correlated gene pairs, indicating that the genes with G4s present at this point result in significantly increased likelihood of coexpression ($P < 0.00001$).

HS3-AUG was located at –16 of the AUG feature and on the coding strand, which was expected to be represented on mRNA and to correspond to 5′UTR. Any G4 found on this spot would be more likely to influence the expression at transcription stage. G4s showed a significant peak at this point along the AUG-aligned profile of osa and the correlated gene pairs correlated to this hotspot were statistically significant ($P < 0.01$), although not as strongly as HS2-TSS and HS4-EXINT.

HS4-EXINT, a significant hotspot located at 24 bases downstream of the first exon–intron boundary, showed a very statistically significant number of correlated gene pairs ($P < 0.00001$), indicating a regulatory role.

Two additional positions were also chosen as negative controls that show neither peaks nor conservation: –1000 of TSS at the coding strand (CS5-CTRL) and +1000 of TTS of the template strand (CS6-CTRL) were chosen due to maximum distance to any feature of the gene and lack of any pronounced peak (Table 2).

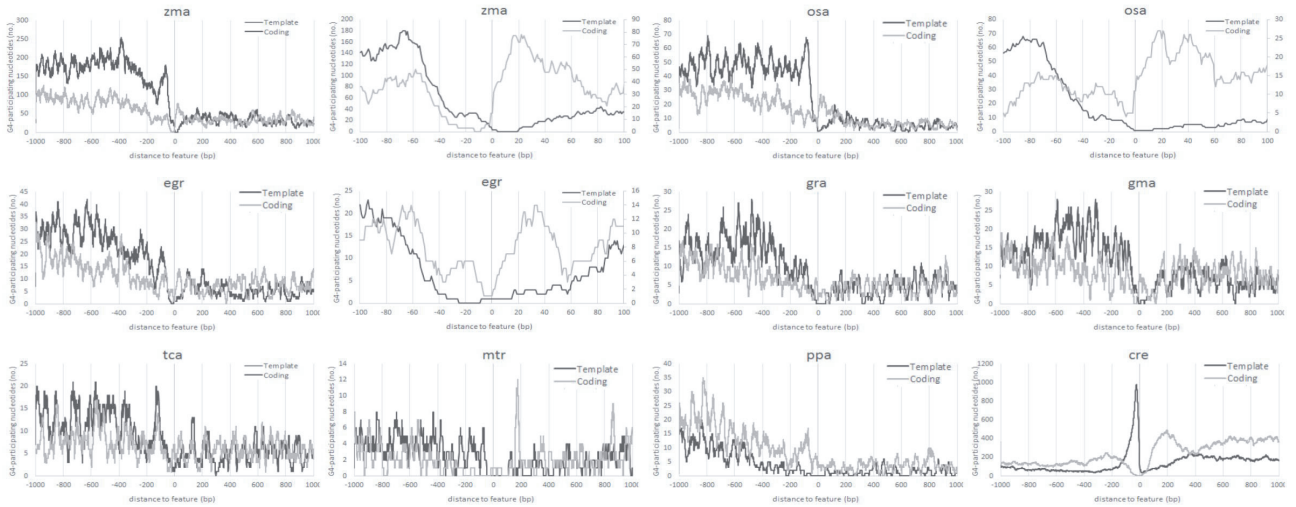

Figure 5. G4-Participating nucleotide distribution profiles in the ±1000 (left) or ±100 (right) nucleotides vicinity of INTEX separately for coding (light gray) and template (dark gray) strands.

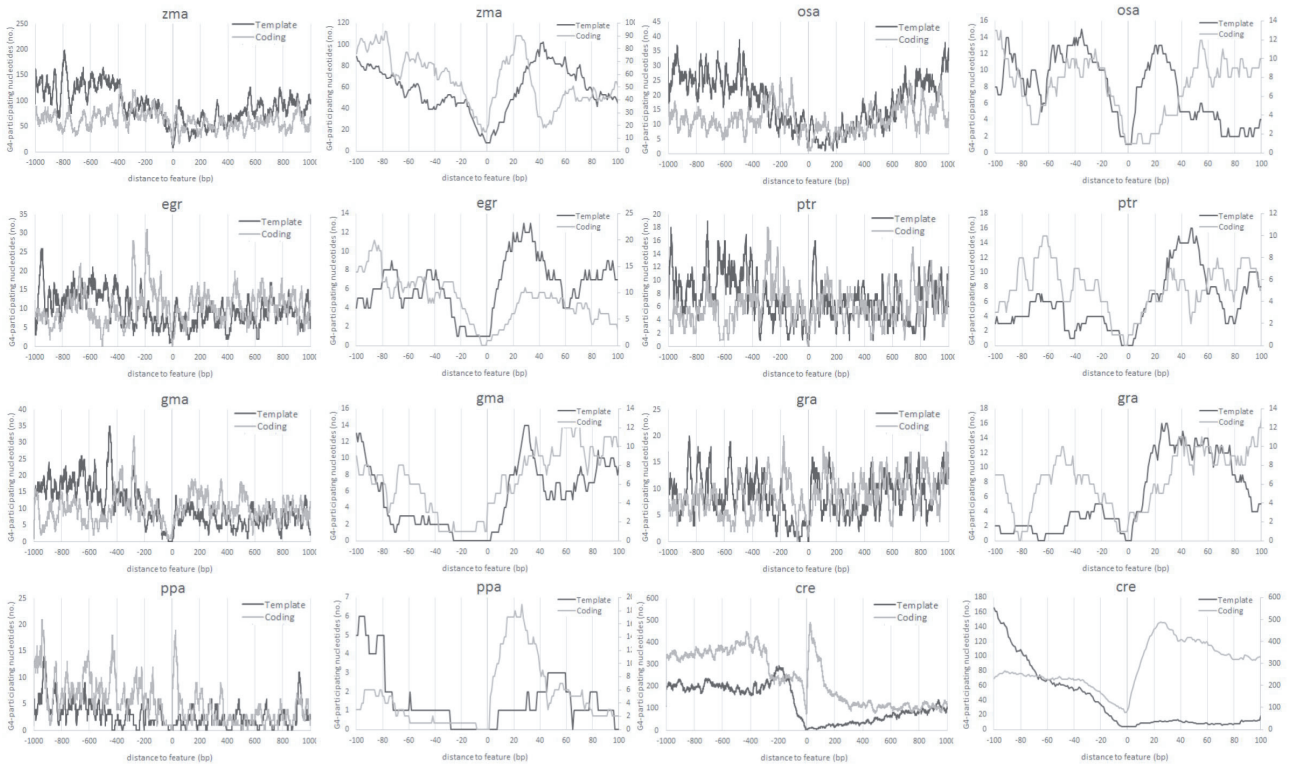

Figure 6. G4-Participating nucleotide distribution profiles in the ±1000 (left) or ±100 (right) nucleotides vicinity of STOP feature separately for coding (light gray) and template (dark gray) strands.

The statistical analysis clearly showed that the negative controls, CS5-CTRL and CS6-CTRL, did not show any significant z-score, indicating that there was no accumulation of correlated genes among genes with G4s at these points as expected, since G4s were not specifically enriched at these points. On the other hand, hotspots that showed specific G4-richness had statistically significant numbers of correlated gene pairs.

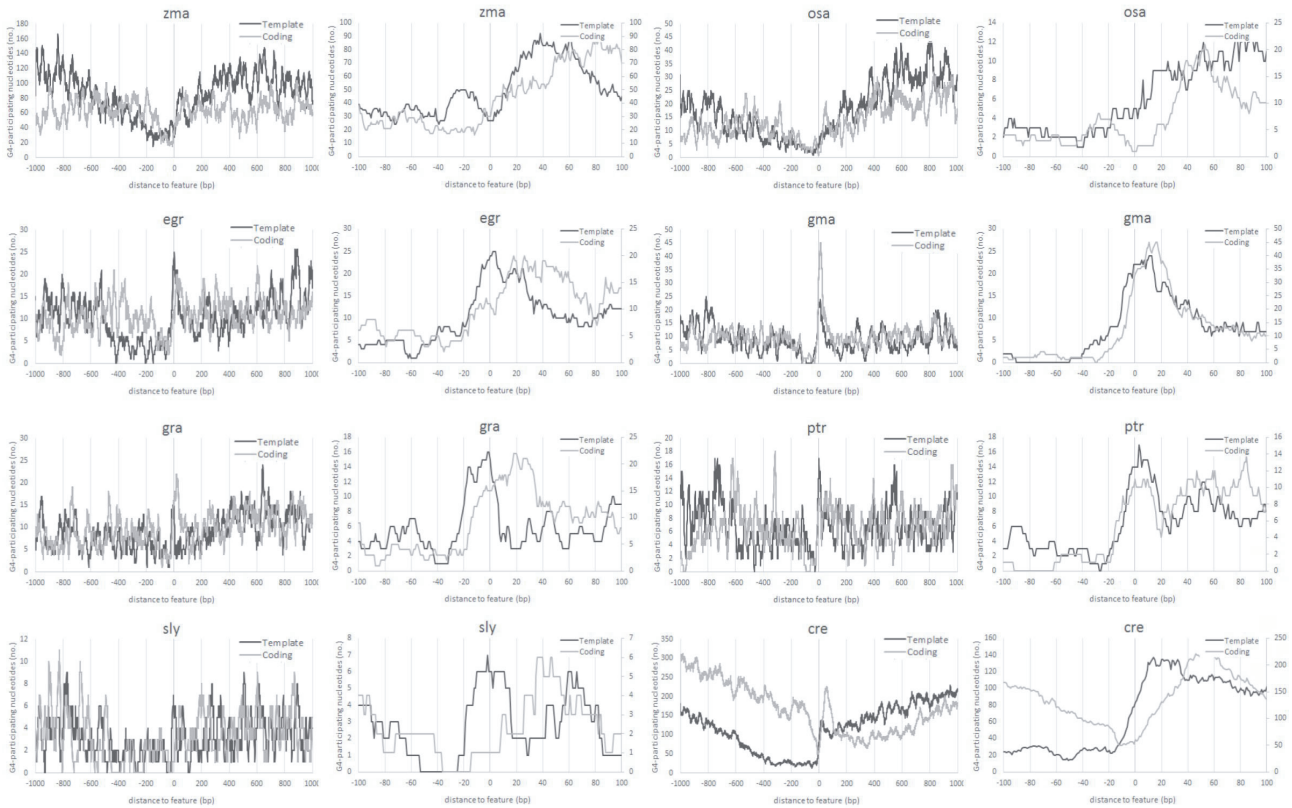

Figure 7. G4-Participating nucleotide distribution profiles in the ±1000 (left) or ±100 (right) nucleotides vicinity of TTS separately for coding (light gray) and template (dark gray) strands.

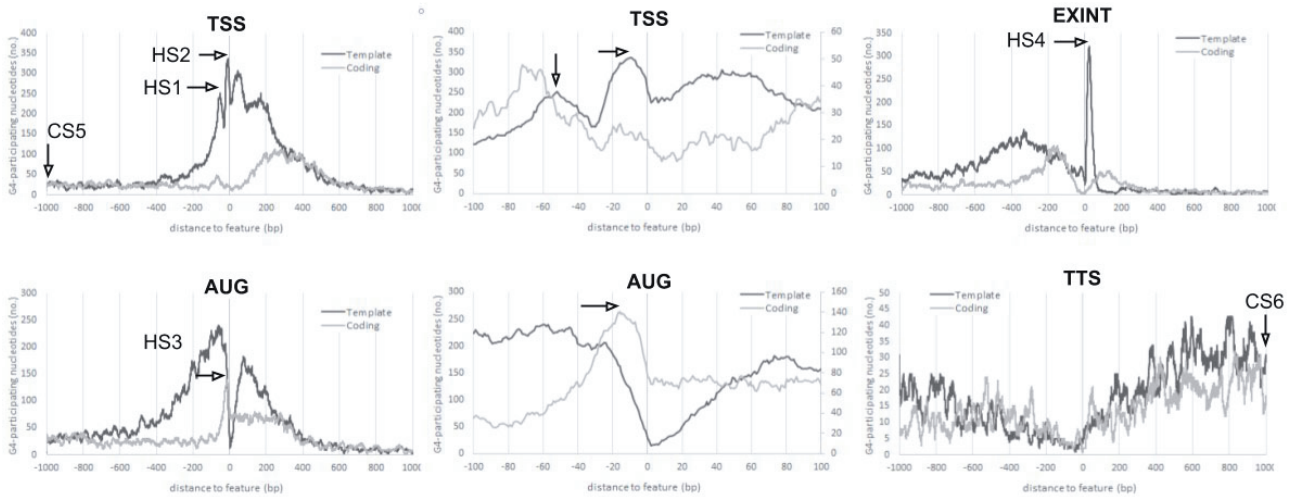

Figure 8. Hotspots and negative controls for *O. sativa* (osa) TSS-, AUG-, and EXINT-aligned profiles selected for analysis.

3.8. G4 identity matrices

As mentioned above, a regulatory element was expected to show distinct topology and be present at a specific location. The hotspots were already predicted to form G4s; however, it is often forgotten that G4 topologies differ from one another and an expanding variety of G4 structures are known. These topologies are dependent on environmental factors, but even more so on the sequence. The positioning of the G-tracts, length of the loops, and even the loop composition add to the final structure. Thus,

it is not completely accurate to regard all these G4s as a single type of molecular element. One type might have stronger interaction with a particular transcription factor and thus result in a stronger coregulation. We previously suggested that G4 topology would have different effects at the molecular level and suggested clustering based on topological similarities (Kaplan et al., 2016). Topologies of the G4s, of course, cannot be identified accurately without extensive analysis of every single G4 computationally predicted in a genome and even that is prone to deformation in vivo. However, we suggested that a practical approach would be the pairwise alignment of the G-quadruplex-forming sequences. Although the loops do not have as strong an influence on the stability of G4 as the G-tracts, they are still influential on its potential interactions as the bases of the loops often look outwards in solution and are presented to the intracellular environment. Thus, an identity matrix was calculated for G4s at each hotspot in order to investigate if gene coregulation is also related to sequences of G4s and, evidently, their topology.

The pairwise alignment scores were calculated for each G4 pair with arbitrarily chosen alignment scores as described in Section 2 and then divided by the length of the shorter G-quadruplex-forming sequence, which equals the maximum score that could have achieved by the pair. The latter step is undertaken to normalize the results between different alignments and provide comparability.

When converted into heat maps, G-quadruplex identity matrices surprisingly did not show any significant similarity in any of the hotspots (Figure 9). In HS3-AUG, an increase in G4 identity within the large cluster suggested that the G4 topology has a marginal connection to the correlation. The high identity G4s shared a common motif: a polyguanine tract of 16 bases. A similar accumulation of G4s with high identity was observed in the smaller cluster of HS2-TSS sharing long polyguanine tracts as their common motif. Interestingly, in HS4-EXINT, the highest G4 identities corresponded to the highest gene expression correlations. These G4s also showed a common polyguanine tract with a minimum length of 15 bases.

4. Discussion

The G4 analysis showed that the most significant shifts or peaks in G4 intensities were observed in the vicinity of the first half of the features, TSS, AUG, and EXINT, where most of the regulatory elements are located. Profiles showed an overall frequency of G4 formation on the template strand in the vicinity of these three features while no significant difference was recorded between template and coding strands for the latter features. Such a difference between template and coding strands, the significant shifts throughout the profiles, and the overall increased density of G4s for TSS, AUG, and EXINT are all indications of the presence of a G4-associated regulatory role in transcription.

One of the most pronounced characteristics in the TSS-aligned profiles is the enriched G4 formation in a broad region in the vicinity of the TSS template, which may be related to transcriptional initiation (Figure 1). Indeed, G4 formation was previously shown to take part in the promoter. Significant formation around the transcription start site may be required for the release of the double-stranded form. This is further supported by the increased coexpression for the genes corresponding to the hotspots upstream of the TSS. Since HS1-TSS and HS2-TSS both are present upstream of the TSS features (Figure 8), the G4s are not represented at mRNA level and their regulatory role is only limited to transcription. Since these hotspots are present on the template strand it is expected that a direct interaction with transcription factors is probable.

For AUG, on the other hand, G4 formation was exclusively on the template strand or the upstream of the coding strand (Figure 3). This was expected since the downstream of AUG corresponds to the first exon and its sequence is mostly determined by the open reading frame of the coding amino acid sequence. A sharp dip in the template strand at the first codon was prominent in most profiles. This strange behavior may be related to avoiding stalling of RNA polymerase at the first codon, which is essential for functional protein production where mutation could not be tolerated. It would be intriguing to see if replacing the first codon within a G4 loop region would influence the expression and could answer why such a particular characteristic was conserved.

Another conserved characteristic peak, referred to as HS3-AUG, was found upstream of AUG on the coding strand, which corresponds to 5′UTR (Figure 8). It is important to note that the gene expression data analyzed here were obtained through whole-exome sequencing (WES) and thus do not represent expression as protein product concentration but rather as mRNA concentration. For that reason, influence of G4s on translation would not be observed. However, our results suggest that 5′UTR G4s may regulate at the mRNA level. The mode of this regulation does not have to be during translation; instead, it may be through regulation of mRNA half-life, a consequence that would be represented in WES. Here we might be observing protection of the 5′-end of mRNAs by G4s through steric hindrance. This effect, in fact, may be put to the test by comparison of the 5′ extension of mRNA sequence reads in raw WES data.

The exon–intron boundary, EXINT, showed a peculiar G4 enrichment on the template strand right after the feature for most species, clearly marking the beginning of the intron (Figure 4). The sharp peak at this location

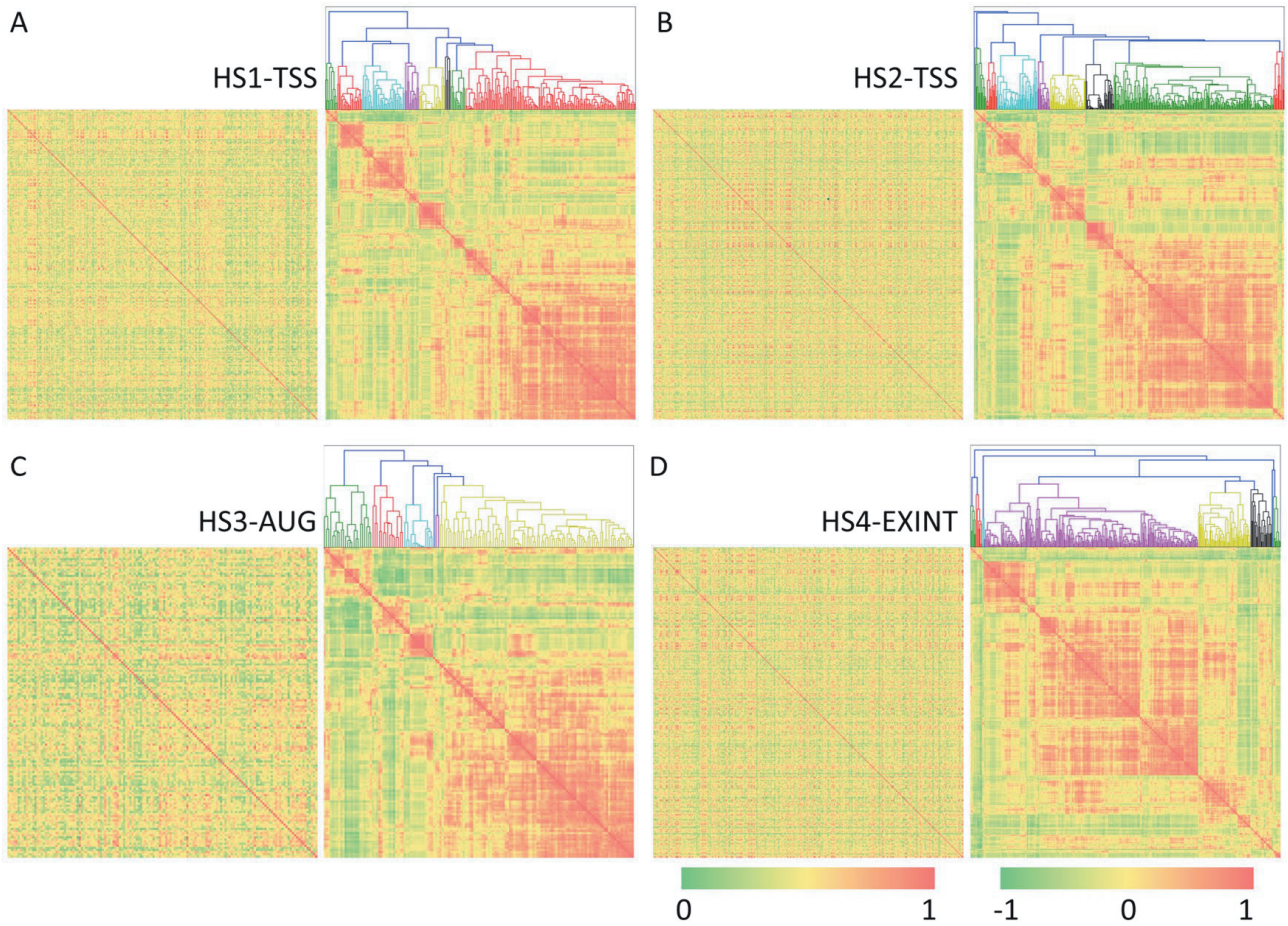

Figure 9. G4 identity matrices (left) of the putative G4-forming sequences and expression similarity matrices (right) of the genes associated to each hotspot. Hierarchical clustering was performed using UPGMA linkage method. Expression similarities were calculated using Pearson correlation and G4 identities were calculated using local pairwise alignment method.

forming HS4-EXINT also indicated coexpression with a high statistical significance that can be explained by a regulatory role (Table 2). However, it should be noted that HS4-EXINT is found on the template strand, which means G4s are formed only at DNA level (Figure 8). Such G4 formation was also recorded for zma in this study and in a previous one (Andorf et al., 2014). It was suggested that this may have a role in stalling RNA polymerase during transcription. This was a rather unexpected result, as it would be simpler to explain a hotspot if it was located on mRNA and possibly taking part in mRNA splicing or premature termination as suggested before (Majewski, 2002). However, DNA-level intronic regulatory elements are not new and were previously shown in plants (Mascarenhas et al. 1990; Hernandez-Garcia and Finer, 2014). Such regulation was only found in the leading intron and is known as intron-mediated enhancement (IME) (Gallegos and Rose, 2015), and, in light of the literature, G4s present in HS4-EXINT are very likely to be IME elements. Such a role of G4s in IME is indeed possible since IME is thought to enhance transcription by making the upstream region of the intron more accessible (Gallegos and Rose, 2015), which coincides with the relation between G-quadruplex formation and the torsion in the flanking duplex DNA (Wang and Lynch, 1996) and the influence of G-quadruplexes on transcription over long distances (Zhang et al., 2013).

It should also be noted that the distribution profiles showed overall resemblance within members of evolutionary branches according to NCBI taxonomy (Figure 2) (Sayers et al., 2009). For members of Poaceae, osa and zma, the distribution profiles showed particularly significant resemblance, which was expected due to evolutionary similarity. Some other species, such as egr, gma, cme, gra, and ptr, also showed similarity to this couple. Evolutionary similarity was also observed among *Arabidopsis*, which, oddly, also showed features similar to sly.

On the other hand, moss and algae species, ppa, olu, and cre, had very different profiles than the rest, especially in the coding strand (Figures 1 and 4). Moving through higher plants, a dramatic increase was observed in the gap between G4 abundance on the template strand and the coding strand, which may be explained as the intracellular mechanisms developing an alternative use for the G4s. These findings indicated that G4 enrichment may be a trait developed through evolution of plants.

When hierarchically clustered similarity matrices were visualized as heat maps, it could clearly be seen that G4-associated genes were not equally correlated among each other within hotspots (Table 2). For instance, in HS1-TSS a large cluster of genes was highly associated among each other and several smaller clusters were present, indicating that not all G4s at this hotspot may have similar coregulatory roles. In fact, the larger cluster indicated that only about half of the genes were associated at this hotspot. Similarly, HS2-TSS and HS2-AUG also showed formation of a large main cluster each, indicating that these genes were indeed coexpressed. Thus, it can be suggested that in most cases when G4 is present there is a high chance that the associated gene expression is correlated. In the case of HS4-EXINT the main cluster formed an even larger cluster, indicating a stronger G4-dependent correlation.

Finally, the evaluation of the G4 identity matrices suggested that most G4s that mediate coexpression did not share topological similarity (Figure 9). This was unexpected since the topology may affect the ability to bind transcription or translation factors. Instead we saw that formation of G4 was adequate for coexpression and did not require specific G4 topology. On the other hand, the constructions of the G4 identity matrices were based on arbitrarily chosen pairwise alignment parameters, and an investigation into more suitable parameters for the alignment of G-quadruplex-forming sequences may provide a better understanding of the relation between the structure and the sequence.

References

Andorf CM, Kopylov M, Dobbs D, Koch KE, Stroupe ME et al. (2014). G-quadruplex (G4) motifs in the maize (*Zea mays* L.) genome are enriched at specific locations in thousands of genes coupled to energy status, hypoxia, low sugar, and nutrient deprivation. Journal of Genetics and Genomics 41 (12): 627–647. doi: 10.1016/j.jgg.2014.10.004

Argout X, Salse J, Aury JM, Guiltinan MJ, Droc G et al. (2011). The genome of *Theobroma cacao*. Nature Genetics 43 (2): 101-108.

Beaudoin JD, Perreault JP (2013). Exploring mRNA 3'-UTR G-quadruplexes: evidence of roles in both alternative polyadenylation and mRNA shortening. Nucleic Acids Research 41 (11): 5898-5911. doi: 10.1093/nar/gkt265

Biswas B, Kandpal M, Jauhari UK, Vivekanandan P (2016). Genome-wide analysis of G-quadruplexes in herpesvirus genomes. BMC Genomics 17 (1): 949. doi: 10.1186/s12864-016-3282-1

Burge S, Parkinson GN, Hazel P, Todd AK, Neidle S (2006). Quadruplex DNA: sequence, topology and structure. Nucleic Acids Resesearch 34 (19): 5402-5415. doi: 10.1093/nar/gkl655

Chan AP, Crabtree J, Zhao Q, Lorenzi H, Orvis J et al. (2010). Draft genome sequence of the oilseed species *Ricinus communis*. Nature Biotechnology 28 (9): 951-956.

Cogoi S, Xodo LE (2006). G-quadruplex formation within the promoter of the KRAS proto-oncogene and its effect on transcription. Nucleic Acids Research 34 (9): 2536-2549. doi: 10.1093/nar/gkl286

Davidson RM, Gowda M, Moghe G, Lin H, Vaillancourt B et al. (2012). Comparative transcriptomics of three *Poaceae* species reveals patterns of gene expression evolution. The Plant Journal 71 (3): 492-502. doi: 10.1111/j.1365-313X.2012.05005.x

Dohm JC, Minoche AE, Holtgräwe D, Capella-Gutiérrez S, Zakrzewski F et al. (2014). The genome of the recently domesticated crop plant sugar beet (*Beta vulgaris*). Nature 505 (7484): 546-549.

Eddy A, Galloway DJ, John DM, Tittley I (1992). Lower plant diversity. In: Groombridge B (editor). Global Biodiversity. Dordrecht, the Netherlands: Springer, pp. 55-63.

Eddy J, Maizels N (2006). Gene function correlates with potential for G4 DNA formation in the human genome. Nucleic Acids Research 34 (14): 3887-3896. doi: 10.1093/nar/gkl529

Fernando H, Reszka AP, Huppert J, Ladame S, Rankin S et al. (2006). A conserved quadruplex motif located in a transcription activation site of the human c-kit oncogene. Biochemistry 45 (25): 7854-7860. doi: 10.1021/bi0601510

Fletcher TM, Sun D, Salazar M, Hurley LH (1998). Effect of DNA secondary structure on human telomerase activity. Biochemistry 37 (16): 5536-5541. doi: 10.1021/bi972681p

Gallegos JE, Rose AB (2015). The enduring mystery of intron-mediated enhancement. Plant Science 237: 8-15. doi: 10.1016/j.plantsci.2015.04.017

Garcia-Mas J, Benjak A, Sanseverino W, Bourgeois M, Mir G et al. (2012). The genome of melon (*Cucumis melo* L.). Proceedings of the National Academy of Sciences of the USA 109 (29): 11872-11877. doi: 10.1073/pnas.1205415109

Garg R, Aggarwal J, Thakkar B (2016). Genome-wide discovery of G-quadruplex forming sequences and their functional relevance in plants. Scientific Reports 6: 28211.

Grand CL, Powell TJ, Nagle RB, Bearss DJ, Tye D et al. (2004). Mutations in the G-quadruplex silencer element and their relationship to c-MYC overexpression, NM23 repression, and therapeutic rescue. Proceedings of the National Academy of Sciences of the USA 101 (16): 6140-6145. doi: 10.1073/pnas.0400460101

He G, Zhu X, Elling AA, Chen L, Wang X et al. (2010). Global epigenetic and transcriptional trends among two rice subspecies and their reciprocal hybrids. The Plant Cell 22 (1): 17-33. doi: 10.1105/tpc.109.072041

Hernandez-Garcia CM, Finer JJ (2014). Identification and validation of promoters and cis-acting regulatory elements. Plant Science 217-218: 109-119. doi: 10.1016/j.plantsci.2013.12.007

Hershman SG, Chen Q, Lee JY, Kozak ML, Yue P et al. (2008). Genomic distribution and functional analyses of potential G-quadruplex-forming sequences in *Saccharomyces cerevisiae*. Nucleic Acids Research 36 (1): 144-156. doi: 10.1093/nar/gkm986

Hu TT, Pattyn P, Bakker EG, Cao J, Cheng JF et al. (2011). The *Arabidopsis lyrata* genome sequence and the basis of rapid genome size change. Nature Genetics 43 (5): 476-481.

Huppert JL (2005). Prevalence of quadruplexes in the human genome. Nucleic Acids Research 33 (9): 2908-2916. doi: 10.1093/nar/gki609

Huppert JL, Balasubramanian S (2007). G-quadruplexes in promoters throughout the human genome. Nucleic Acids Research 35 (2): 406-413. doi: 10.1093/nar/gkl1057

Huppert JL, Bugaut A, Kumari S, Balasubramanian S (2008). G-quadruplexes: the beginning and end of UTRs. Nucleic Acids Research 36 (19): 6260-6268. doi: 10.1093/nar/gkn511

International Rice Genome Sequencing Project. (2005). The map-based sequence of the rice genome. Nature 436 (7052): 793-800.

Jaillon O, Aury JM, Noel B, Policriti A, Clepet C et al. (2007). The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. Nature 449 (7161): 463-467.

Kaplan OI, Berber B, Hekim N, Doluca O (2016). G-quadruplex prediction in *E. coli* genome reveals a conserved putative G-quadruplex-hairpin-duplex switch. Nucleic Acids Research 44 (19): 9083-9095. doi: 10.1093/nar/gkw769

Kawahara Y, de la Bastide M, Hamilton JP, Kanamori H, McCombie WR et al. (2013). Improvement of the *Oryza sativa* Nipponbare reference genome using next generation sequence and optical map data. Rice 6 (1): 4. doi: 10.1186/1939-8433-6-4

Kwok CK, Ding Y, Shahid S, Assmann SM, Bevilacqua PC (2015). A stable RNA G-quadruplex within the 5'-UTR of *Arabidopsis thaliana* ATR mRNA inhibits translation. Biochemical Journal 467 (1): 91-102. doi: 10.1042/BJ20141063

Majewski J (2002). Distribution and characterization of regulatory elements in the human genome. Genome Research 12 (12): 1827-1836. doi: 10.1101/gr.606402

Mascarenhas D, Mettler IJ, Pierce DA, Lowe HW (1990). Intron-mediated enhancement of heterologous gene expression in maize. Plant Molecular Biology 15 (6): 913-920. doi: 10.1007/BF00039430

Merchant SS, Prochnik SE, Vallon O, Harris EH, Karpowicz SJ et al. (2007). The *Chlamydomonas* genome reveals the evolution of key animal and plant functions. Science 318 (5848): 245-250. doi: 10.1126/science.1143609

Michener CD, Sokal RR. (1957). A quantitative approach to a problem of classification. Evolution 11: 490-499.

Mullen MA, Olson KJ, Dallaire P, Major F, Assmann SM et al. (2010). RNA G-Quadruplexes in the model plant species *Arabidopsis thaliana*: prevalence and possible functional roles. Nucleic Acids Research 38 (22): 8149-8163. doi: 10.1093/nar/gkq804

Myburg AA, Grattapaglia D, Tuskan GA, Hellsten U, Hayes RD et al. (2014). The genome of *Eucalyptus grandis*. Nature 510 (7505): 356-362.

Palenik B, Grimwood J, Aerts A, Rouzé P, Salamov A et al. (2007). The tiny eukaryote *Ostreococcus* provides genomic insights into the paradox of plankton speciation. Proceedings of the National Academy of Sciences of the USA 104 (18): 7705-7710. doi: 10.1073/pnas.0611046104

Prochnik S, Marri PR, Desany B, Rabinowicz PD, Kodira C et al. (2012). The *Cassava* genome: current progress, future directions. Tropical Plant Biology 5 (1): 88-94. doi: 10.1007/s12042-011-9088-z

Proost S, Van Bel M, Vaneechoutte D, Van de Peer Y, Inze D et al. (2014). PLAZA 3.0: an access point for plant comparative genomics. Nucleic Acids Research 43 (D1): D974-D981.

Rensing SA, Lang D, Zimmer AD, Terry A, Salamov A et al. (2008). The *Physcomitrella* genome reveals evolutionary insights into the conquest of land by plants. Science 319 (5859): 64-69.

Sayers EW, Barrett T, Benson DA, Bryant SH, Canese K et al. (2009). Database resources of the National Center for Biotechnology Information. Nucleic Acids Research 37 (Database issue): D5-15.

Schmutz J, Cannon SB, Schlueter J, Ma J, Mitros T et al. (2010). Genome sequence of the palaeopolyploid soybean. Nature 463 (7278): 178-183.

Schnable PS, Ware D, Fulton RS, Stein JC, Wei F et al. (2009). The B73 maize genome: complexity, diversity, and dynamics. Science 326 (5956): 1112-1115. doi: 10.1126/science.1178534

Slotte T, Hazzouri KM, Ågren JA, Koenig D, Maumus F et al. (2013). The *Capsella rubella* genome and the genomic consequences of rapid mating system evolution. Nature Genetics 45 (7): 831-835.

Sun D, Liu WJ, Guo K, Rusche JJ, Ebbinghaus S et al. (2008). The proximal promoter region of the human vascular endothelial growth factor gene has a G-quadruplex structure that can be targeted by G-quadruplex-interactive agents. Molecular Cancer Therapeutics 7 (4): 880-889. doi: 10.1158/1535-7163.MCT-07-2119

Takahashi H, Nakagawa A, Kojima S, Takahashi A, Cha BY et al. (2012). Discovery of novel rules for G-quadruplex-forming sequences in plants by using bioinformatics methods. Journal of Bioscience and Bioengineering 114 (5): 570-575. doi: 10.1016/j.jbiosc.2012.05.017

The *Arabidopsis* Genome Initiative (2000). Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. Nature 408: 796-815.

Todd AK, Johnston M, Neidle S (2005). Highly prevalent putative quadruplex sequence motifs in human DNA. Nucleic Acids Research 33 (9): 2901-2907. doi: 10.1093/nar/gki553

Tomato Genome Consortium (2012). The tomato genome sequence provides insights into fleshy fruit evolution. Nature 485 (7400): 635-641.

Tuskan GA, Difazio S, Jansson S, Bohlmann J, Grigoriev I et al. (2006). The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray). Science 313 (5793): 1596-1604. doi: 10.1126/science.1128691

Verde I, Abbott AG, Scalabrin S, Jung S, Shu S et al. (2013). The high-quality draft genome of peach (*Prunus persica*) identifies unique patterns of genetic diversity, domestication and genome evolution. Nature Genetics 45 (5): 487-494. doi: 10.1038/ng.2586

Verma A, Halder K, Halder R, Yadav VK, Rawal P et al. (2008). Genome-wide computational and expression analyses reveal G-quadruplex DNA motifs as conserved cis-regulatory elements in human and related species. Journal of Medicinal Chemistry 51 (18): 5641-5649. doi: 10.1021/jm800448a

Wang JC, Lynch SA (1996). Effects of DNA supercoiling on gene expression. In: Lin ECC, Lynch SA (editors). Regulation of Gene Expression in *Escherichia coli*. Boston, MA, USA: Springer, pp. 127-147.

Wang K, Wang Z, Li F, Ye W, Wang J et al. (2012). The draft genome of a diploid cotton *Gossypium raimondii*. Nature Genetics 44 (10): 1098-1103.

Wang Y, Zhao M, Zhang Q, Zhu GF, Li FF et al. (2015). Genomic distribution and possible functional roles of putative G-quadruplex motifs in two subspecies of *Oryza sativa*. Computational Biology and Chemistry 56: 122-130. doi: 10.1016/j.compbiolchem.2015.04.009

Wieland M, Hartig JS (2009). Investigation of mRNA quadruplex formation in *Escherichia coli*. Nature Protocols 4 (11): 1632-1640. doi: 10.1038/nprot.2009.111

Xu Q, Chen LL, Ruan X, Chen D, Zhu A et al. (2013). The draft genome of sweet orange (*Citrus sinensis*). Nature Genetics 45 (1): 59-66.

Young ND, Debellé F, Oldroyd GED, Geurts R, Cannon SB et al. (2011). The *Medicago* genome provides insight into the evolution of rhizobial symbioses. Nature 480 (7378): 520-524.

Zemach A, McDaniel IE, Silva P, Zilberman D (2010). Genome-wide evolutionary analysis of eukaryotic DNA methylation. Science 328 (5980): 916-919. doi: 10.1126/science.1186366

Zhang C, Liu HH, Zheng KW, Hao YH, Tan Z (2013). DNA G-quadruplex formation in response to remote downstream transcription activity: long-range sensing and signal transducing in DNA double helix. Nucleic Acids Research 41 (14): 7144-7152. doi: 10.1093/nar/gkt443

Zhao Y, Du Z, Li N (2007). Extensive selection for the enrichment of G4 DNA motifs in transcriptional regulatory regions of warm blooded animals. FEBS Letters 581 (10): 1951-1956. doi: 10.1016/j.febslet.2007.04.017