Genome-wide analysis reveals regulatory role of G4 DNA in gene transcription

Zhuo Du, Yiqiang Zhao and Ning Li

Genome Res. 2008 18: 233-241 originally published online December 20, 2007
Access the most recent version at doi:10.1101/gr.6905408

Letter

Genome-wide analysis reveals regulatory role of G4 DNA in gene transcription

Zhuo Du,[1] Yiqiang Zhao,[1] and Ning Li[2]

State Key Laboratory for Agrobiotechnology, China Agricultural University, Beijing, 10094, People's Republic of China

G-quadruplex or G4 DNA, a four-stranded DNA structure formed in G-rich sequences, has been hypothesized to be a structural motif involved in gene regulation. In this study, we examined the regulatory role of potential G4 DNA motifs (PG4Ms) located in the putative transcriptional regulatory region (TRR, −500 to +500) of genes across the human genome. We found that PG4Ms in the 500-bp region downstream of the annotated transcription start site (TSS; $PG4M_{D500}$) are associated with gene expression. Generally, $PG4M_{D500}$-positive genes are expressed at higher levels than $PG4M_{D500}$-negative genes, and an increased number of $PG4M_{D500}$ provides a cumulative effect. This observation was validated by controlling for attributes, including gene family, function, and promoter similarity. We also observed an asymmetric pattern of $PG4M_{D500}$ distribution between strands, whereby the frequency of $PG4M_{D500}$ in the coding strand is generally higher than that in the template strand. Further analysis showed that the presence of $PG4M_{D500}$ and its strand asymmetry are associated with significant enrichment of RNAP II at the putative TRR. On the basis of these results, we propose a model of G4 DNA-mediated stimulation of transcription with the hypothesis that $PG4M_{D500}$ contributes to gene transcription by maintaining the DNA in an open conformation, while the asymmetric distribution of $PG4M_{D500}$ considerably reduces the probability of blocking the progression of the RNA polymerase complex on the template strand. Our findings provide a comprehensive view of the regulatory function of G4 DNA in gene transcription.

[Supplemental material is available online at www.genome.org.]

Genomic DNA predominantly exists in the double-stranded conformation throughout most of the cell cycle; however, certain guanine-rich sequences can fold spontaneously into a four-stranded DNA structure known as a G-quadruplex or G4 DNA. The structure of G4 DNA, which comprises stacked G-tetrads, a square planar arrangement of four guanine bases stabilized by Hoogsteen GG pairing, is extremely stable under physiological conditions (Gellert et al. 1962; Guschlbauer et al. 1990; Han and Hurley 2000; Keniry 2000; Shafer and Smirnov 2000; Simonsson 2001; Burge et al. 2006).

Although relatively little is known about the detailed molecular mechanism by which G4 DNA influences genome function, local DNA structure alternative to the double-stranded conformation might provide regulatory motifs important for gene regulation. Recent studies have shown that the G4 DNA structures formed in the regulatory regions can regulate gene expression (Han and Hurley 2000; Dexheimer et al. 2006; Maizels 2006; Rawal et al. 2006; Fry 2007). A well-known example is the repressive effect of G4 DNA on transcription of the human *MYC* gene. The transcriptional activity of the *MYC* gene is reduced considerably when a parallel G4 DNA, formed in the nuclease-hypersensitive element III_1 upstream of the P1 promoter, is stabilized by the G4 ligand TMPyP4. In comparison, a G4 DNA-disrupting mutation caused a threefold increase in the basal promoter activity (Grand et al. 2002; Siddiqui-Jain et al. 2002; Seenisamy et al. 2004; Ambrus et al. 2005). Similarly, a reporter assay indicated that stabilization of G4 DNA in the nuclease-hypersensitive polypurine–polypyrimidine element of the *KRAS* promoter caused an 80% decrease in transcriptional activity (Cogoi and Xodo 2006). In contrast, several studies have shown a stimulatory role of G4 DNA motifs in gene expression. For example, biochemical and biophysical analyses have indicated that the formation of G4 DNAs in the regulatory regions of the chicken beta-globin and human insulin genes activates transcription through the binding of the G4 DNA-specific proteins (Lewis et al. 1988; Clark et al. 1990; Kennedy and Rutter 1992; Catasti et al. 1996; Lew et al. 2000; Dexheimer et al. 2006). The G-rich nontemplate strand of rDNA has also been reported to be capable of forming G4 DNA, and this structure has been hypothesized to contribute to the high efficiency of rRNA transcription by interacting with a G4 DNA-specific binding protein, nucleolin (Hanakahi et al. 1999; Maizels 2006).

In addition to these sporadic examples, subsequent studies identified the existence of G4 DNA-forming sequences in the regulatory regions of many other genes. Particularly, G4 DNA motifs have been identified in the promoters of *KIT* (Rankin et al. 2005; Fernando et al. 2006), *HIF1A* (De Armond et al. 2005), *VEGFA* (Sun et al. 2005), *BCL2* (Dai et al. 2006), *RB1* (Xu and Sugiyama 2006), and several muscle-specific genes (Yafe et al. 2005). Genome-wide searches have revealed that potential G4 DNA motifs (PG4Ms) are highly prevalent in the genome and are overrepresented in promoters (Huppert and Balasubramanian 2005, 2007; Todd et al. 2005; Rawal et al. 2006; Du et al. 2007). Moreover, a comparative analysis of PG4Ms across animal species has shown that they are strongly enriched in the transcriptional regulatory region (TRR, defined as 500 bp upstream and downstream of the transcription start site) in warm-blooded animals (Zhao et al. 2007). These combined findings lead to the hypothesis that the G4 DNA structures may be common regulatory elements that play roles in transcriptional regulation through structural transitions in DNA.

[1]These authors contributed equally to this work.
[2]Corresponding author.
E-mail ninglbau@public3.bta.net.cn; fax 86-10-62733904.
Article published online before print. Article and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.6905408.

Although enrichment of PG4Ms in the putative regulatory region has been identified by computational approaches and their role in gene expression has been confirmed experimentally in some cases, a global analysis of the relationship between PG4Ms and gene expression under physiological conditions has not yet been conducted. Is G4 DNA-mediated transcriptional regulation a general feature of gene regulation in the human genome? Here, we address this issue by using microarrary expression data. We found that PG4Ms increase gene transcription and propose a model of G4 DNA-mediated stimulation of transcription.

Results

$PG4M_{D500}$ is associated with the expression levels of human genes

The region of the genome proximal to the TSS is essential for the regulation of transcription. PG4Ms have been reported to be extensively enriched in the putative TRR (1-kb window around the annotated TSS, TSS ± 500 bp) (Zhao et al. 2007), supporting the hypothesis that PG4Ms act as transcriptional regulatory elements. If this is true, an association between PG4Ms and gene expression would be expected. To test this hypothesis, we investigated the influence of PG4Ms in the putative TRR ($PG4M_{TRR}$) on gene expression by performing a computational analysis on a data set of 13,276 nonredundant human RefSeq genes with known expression values. Using the Quadparser program (Huppert and Balasubramanian 2005), we identified a total of 17,779 $PG4M_{TRR}$ in 8214 of the 13,276 genes. We found that the frequency of $PG4M_{TRR}$ (number of PG4Ms per kilobase) was positively correlated with the mean expression level averaged across 79 human tissues/cells (Spearman $\rho = 0.474$, $P < 0.001$).

To further examine this relationship, we grouped the human genes according to the number of $PG4M_{TRR}$ they contained (0, 1, 2, 3, and ≥4), and then compared the mean expression levels (log2-transformed) among these groups using ANOVA. As expected, the mean expression level of $PG4M_{TRR}$-positive genes (genes containing one or more $PG4M_{TRR}$) was significantly higher than that of $PG4M_{TRR}$-negative genes (Table 1).

The frequency of PG4Ms has previously been shown to exhibit a bimodal distribution in the TSS-flanking regions, with the first peak in the 500-bp region upstream of the TSS (–500 to TSS, U500) and the second peak in the 500-bp region downstream of the TSS (TSS to +500, D500) (Zhao et al. 2007). Thus, we next analyzed these two regions separately. We found that the positive correlation between the frequency of PG4Ms and gene expression no longer existed for the U500 region, but remained significant for the D500 region (Spearman $\rho = 0.594$, $P < 0.001$) (Fig. 1A,B). When ANOVA was restricted to $PG4M_{U500}$, the differences in the mean expression levels became ambiguous among groups of genes with different numbers of $PG4M_{U500}$ (0, 1, 2, and ≥3) (Table 1). In the D500 region, not only did PG4M-positive genes continue to be expressed at a significantly higher level than PG4M-negative genes, but the expression level, intuitively, increased with the number of $PG4M_{D500}$ (Table 1; the lack of a significant difference in expression between the groups with higher numbers of $PG4M_{D500}$ and those with lower numbers might have been due to the insufficient number of genes). To confirm the effect of the number of $PG4M_{D500}$ on gene expression, we repeated the correlation analysis on the $PG4M_{D500}$-positive genes alone. Again, this resulted in a significantly positive correlation (Spearman $\rho = 0.271$, $P = 0.042$), validating the cumulative effect. Together, these results indicate that $PG4M_{D500}$, but not $PG4M_{U500}$, correlates with gene expression: $PG4M_{D500}$-positive genes tend to be expressed at a higher level than $PG4M_{D500}$-negative genes, and the number of $PG4M_{D500}$ shows a cumulative effect on the expression levels.

Taking into account the expression variation among tissues, we also performed a multivariate analysis of variance to test for differences in gene expression levels among $PG4M_{D500}$ groups. Again, the result was that $PG4M_{D500}$-positive genes were expressed at a significantly higher level than that of $PG4M_{D500}$-negative genes (Pillai's Trace, F = 3.44, numerator df = 237, denominator df = 39,588, $P < 0.001$). Detailed investigations of gene expression for each tissue/cell type showed that gene expression levels were higher for $PG4M_{D500}$-positive than those for $PG4M_{D500}$-negative genes in all of the 79 human tissues/cells (Fig. 2), of which 76 exhibited a statistically significant difference (Supplemental Table S1). In these 76 tissues/cells, the median gene expression levels (untransformed data) for $PG4M_{D500}$-positive genes were 5.5%–43.9% higher than those for $PG4M_{D500}$-negative genes, with an average increase of 22.71%

Table 1. Comparison of the expression levels among groups of genes containing different numbers of $PG4M_{D500}$

Location	Number of PG4Ms	Number of genes	Mean expression level (log2-transformed)	*F*-value	*P*-value	Significant in post hoc comparisons
TRR	0	5062	8.674	11.37	<0.001	0–1
	1	3442	8.839			0–2
	2	2205	8.864			0–3
	3	1326	8.830			0–≥4
	≥4	1241	8.857			
U500 region	0	7665	8.763	1.93	0.122	/
	1	3393	8.831			
	2	1403	8.752			
	≥3	815	8.791			
D500 region	0	7556	8.697	20.6	<0.001	0–1
	1	3581	8.872			0–2
	2	1463	8.913			0–≥3
	≥3	676	8.957			
Coding strand of the D500 region	0	9261	8.725	23.5	<0.001	0–1
	1	2946	8.900			0–≥2
	≥2	1069	8.941			
Template strand of the D500 region	0	10641	8.757	7.47	<0.001	0–1
	1	2053	8.867			0–≥2
	≥2	582	8.914			

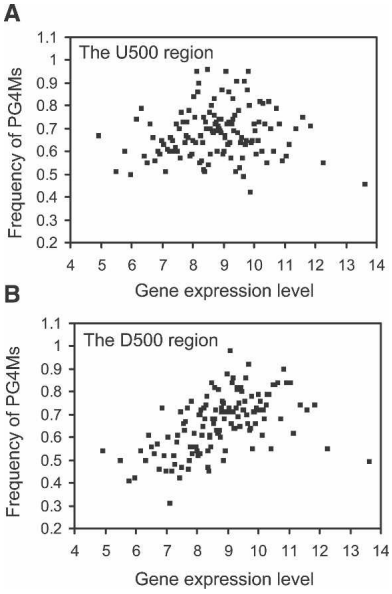

Figure 1. Relationship between PG4Ms and gene expression level. Changes in the frequency of $PG4M_{U500}$ (*A*) and $PG4M_{D500}$ (*B*) with the increased gene expression level are plotted. Genes are ranked by expression level, with each point representing the mean expression level and the frequency of PG4Ms calculated for every 100 genes.

(Supplemental Table S1). Lung, whole brain, dendritic cells, 721 B lymphoblasts, and colorectal adenocarcinoma were the five tissues/cells that showed the greatest expression increases (>35%). These results indicate that the effect of $PG4M_{D500}$ on gene expression is apparently ubiquitous across human tissues/cells.

$PG4M_{D500}$ is associated with gene expression after controlling for gene attributes

Our results, described above, suggest that $PG4M_{D500}$ affects gene expression. Specifically, $PG4M_{D500}$-positive genes are expressed at a higher level than $PG4M_{D500}$-negative genes. Because gene expression is a highly integrated and coordinated process controlled by many factors, to isolate the effect of $PG4M_{D500}$ and to validate our results, we further controlled for gene attributes (gene family, function, and promoter similarity) that potentially influence gene expression and compared the expression levels between $PG4M_{D500}$-negative and $PG4M_{D500}$-positive genes.

We compared the expression of the genes belonging to the same gene family, but carrying different numbers of $PG4M_{D500}$. Only gene families that contained 10 or more members were selected for study. In 52 of the 77 gene families tested, expression levels of $PG4M_{D500}$-positive genes was generally higher than that of $PG4M_{D500}$-negative genes (the squares are generally below the angle bisector; $\chi^2 = 9.468$, df = 1, $P = 0.002$) (Fig. 3A; Supplemental Table S2), consistent with the notion that the presence of $PG4M_{D500}$ increases gene expression. Similar results were obtained when the gene function was controlled: 50 of the 59 categories showed greater gene expression for $PG4M_{D500}$-positive genes than $PG4M_{D500}$-negative genes ($\chi^2 = 28.492$, df = 1, $P < 0.001$) (Fig. 3B; Supplemental Table S3).

More importantly, given that the activity of promoters significantly influences gene expression levels, we normalized the influence of the promoters and measured the expression level of those genes that carried different numbers of $PG4M_{D500}$. An assumption was made that promoters exhibiting a high sequence identity would have similar transcriptional activity. Using reciprocal BLAST, we identified 15 promoter clusters (Supplemental Table S4) that exhibited a considerable sequence similarity (score ≥50 and match length ≥40) in the core promoter region (defined

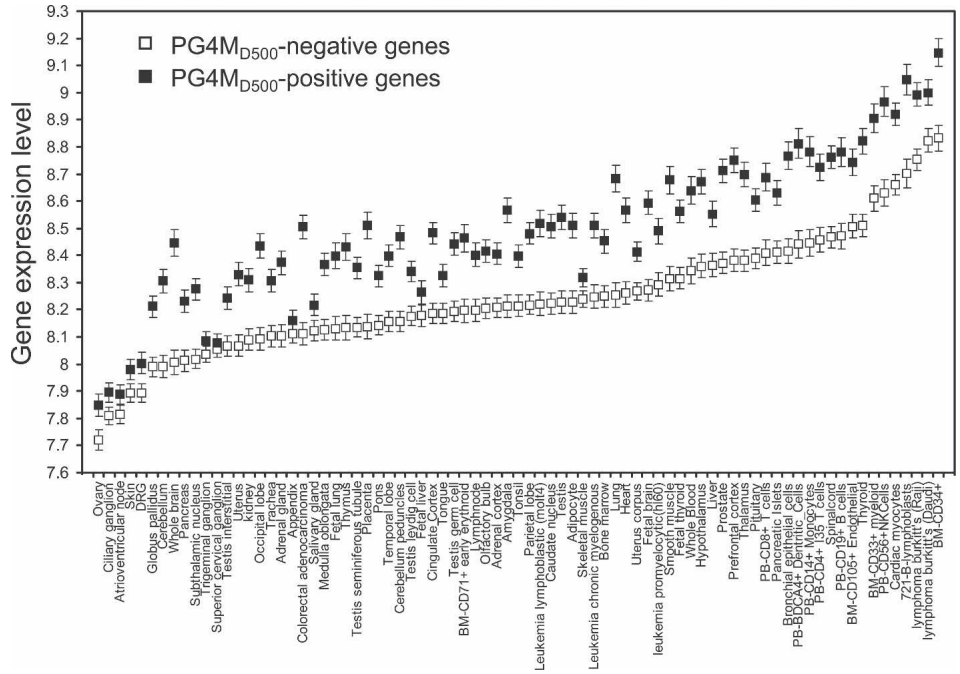

Figure 2. Influence of $PG4M_{D500}$ on gene expression for each tissue/cell. Comparison of the gene expression levels between $PG4M_{D500}$-positive (black squares) and $PG4M_{D500}$-negative genes (open squares) in each human tissue/cell type. Error bars represent the 95% confidence interval of the mean expression level.

as the –100 to +50 region) but carried different numbers of PG4Ms in the D500 region. To ensure reliability of the data and to provide sufficient information, we considered neither gene clusters that contained fewer than three members nor clusters that contained more than six members if there was only a single $PG4M_{D500}$-positive or $PG4M_{D500}$-negative gene. In each cluster, we compared the expression level between genes with and without $PG4M_{D500}$. We found that for 73.33% (11 of 15) of the clusters, the expression level of $PG4M_{D500}$-positive genes was higher than of $PG4M_{D500}$-negative genes (Supplemental Table S4), consistent with the results obtained from analysis of the full data set.

One promoter cluster, whose members belong to the same gene family, was chosen for detailed analysis (Supplemental Fig. S1). This cluster contains three members (*LILRA3, LILRB3*, and *LILRB4*) of the leukocyte immunoglobulin-like receptor (LILR), a family of immunoreceptors that are expressed in a range of immune cells and mediate diverse roles in immune regulation (Fanger et al. 1999). Multiple sequence alignment showed that these genes possess a high degree of sequence similarity in the putative TRR, especially the putative core promoter, but they contain different numbers of PG4Ms in the D500 region. As these genes are members of the same family and possess high similarity in their promoter sequences, we proposed that the significant differences in their expression levels could not be explained simply by different means of gene regulation or promoter activities. The average expression level of $PG4M_{D500}$-positive gene *LILRB3* (8.930) is higher than that of $PG4M_{D500}$-negative genes, *LILRA3* (8.271) and *LILRB4* (8.339).

Together, these results show that even after controlling for gene attributes that potentially influence gene expression, the influence of $PG4M_{D500}$ on gene expression remains significant.

Strand asymmetry of $PG4M_{D500}$

A detailed comparison of PG4Ms between coding ($PG4M_{cod}$) and template ($PG4M_{tem}$) strands in the 10-kb TSS-flanking region revealed a significant asymmetric pattern of distribution. Specifically, the frequency of $PG4M_{cod}$ was notably higher than that of $PG4M_{tem}$ in an ~1.5-kb region downstream of the TSS. The maximal difference was observed in the D500 region, where the frequency of $PG4M_{cod}$ was 60.39% higher than that of $PG4M_{tem}$ (Fig. 4A). However, this strand asymmetry was not observed in either the U500 region or the rest of the transcripts (data not shown); it was also not observed in the human pseudogenes, which are not subjected to evolutionary constraints since they are inactive (Fig. 4A). To exclude the possibility that a few genes with extreme values were responsible for this difference, a bootstrap analysis was performed for subsets containing 2000 genes. All 1000 replications showed a higher frequency of $PG4M_{cod}$ than $PG4M_{tem}$, suggesting that the difference was not due to an effect of relatively few genes. We also classified the genes according to the following criteria in the D500 regions:

(1) $PG4M_{cod}$-positive and $PG4M_{tem}$-negative ($PG4M_{cod+tem-}$);
(2) $PG4M_{cod}$-negative and $PG4M_{tem}$-positive ($PG4M_{cod-tem+}$);
(3) more $PG4M_{cod}$ than $PG4M_{tem}$ ($PG4M_{cod>tem}$); and
(4) fewer $PG4M_{cod}$ than $PG4M_{tem}$ ($PG4M_{cod<tem}$).

We found that the number of $PG4M_{cod+tem-}$ ($n = 3085$) and $PG4M_{cod>tem}$ genes ($n = 3298$) was significantly greater than the corresponding number of $PG4M_{cod-tem+}$ ($n = 1705$) and $PG4M_{cod<tem}$ genes ($n = 1854$), respectively, when considering $PG4M_{D500}$ ($\chi^2 = 485.089$, df = 1, and $\chi^2 = 502.160$, df = 1, $P < 0.001$ in both cases) (Fig. 4B, black squares). However, no significant difference was found when the U500 region was examined (Fig. 4B, gray squares). Furthermore, analysis of the strand distribution of PG4Ms in five warm-blooded animals (chimpanzee, rat, mouse, cow, and chicken) demonstrated that this strand asymmetry of $PG4M_{D500}$ was evolutionarily conserved (Supplemental Fig. S2).

The GC content of the D500 region also had an asymmetric distribution, in that there was a general excess of G over C in the coding strand of most human genes. We therefore measured the asymmetry of PG4Ms and GC as $A_{PG4M} = (PG4M_{cod} - PG4M_{tem})/(PG4M_{cod} + PG4M_{tem})$ and $A_{GC} = (G - C)/(G + C)$, respectively, and found that these two variables were highly correlated (Spearman $\rho = 0.618$, $P < 0.001$). This raises a question as to whether the asymmetry of $PG4M_{D500}$ is caused by the asymmetry in the GC content or vice versa. To address this question, we masked all $PG4M_{D500}$ sequences and recalculated the correlation coefficient. The G and C within the $PG4M_{D500}$ accounted for only 5.8% and 4.8% of the total G and C of the D500 region, respectively; in the D500 region, if the asymmetry of PG4Ms is an embodiment of the background asymmetry of the GC content, we would expect the correlation coefficient to be almost equal to that calculated from the unmasked data. However, there was a sharp decrease in the correlation coefficient, from 0.618 to 0.344, when the $PG4M_{D500}$ sequences were masked. This suggests that the asymmetry of PG4Ms is not a byproduct of the background asymmetry in the GC content, but the asymmetry of GC observed in the D500 region is, to some extent, due to the asymmetry of PG4M. Considering these facts together, we conclude that the asymmetry of PG4Ms is an intrinsic characteristic that has been favored by natural selection in the D500 region.

The strand asymmetry of $PG4M_{D500}$ also leads to the following questions: (1) whether $PG4M_{D500}$ in both strands influences gene expression, and (2) whether the asymmetry of $PG4M_{D500}$ affects gene expression. To address these issues, we distinguished $PG4M_{D500}$ ac-

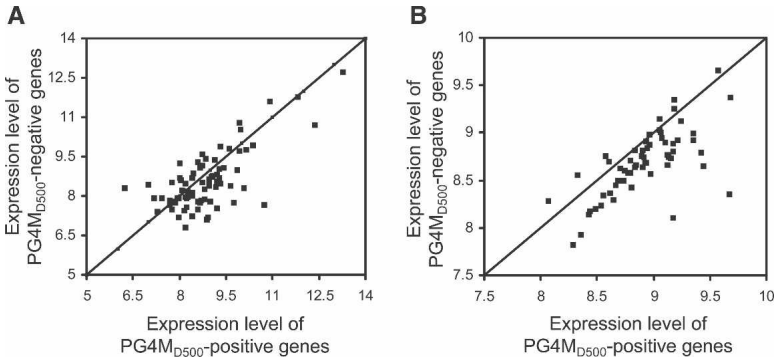

Figure 3. Comparison of expression levels between $PG4M_{D500}$-positive and $PG4M_{D500}$-negative genes when gene family and function are controlled. Human genes are clustered according to gene family (*A*) and gene function (*B*), and mean expression levels are compared between $PG4M_{D500}$-positive (*X*-axis) and $PG4M_{D500}$-negative genes (*Y*-axis) in each cluster. The angle bisector represents equal expression levels. The squares *below* the angle bisector indicate that the expression level of $PG4M_{D500}$-positive genes is higher than that of $PG4M_{D500}$-negative genes, and vice versa.

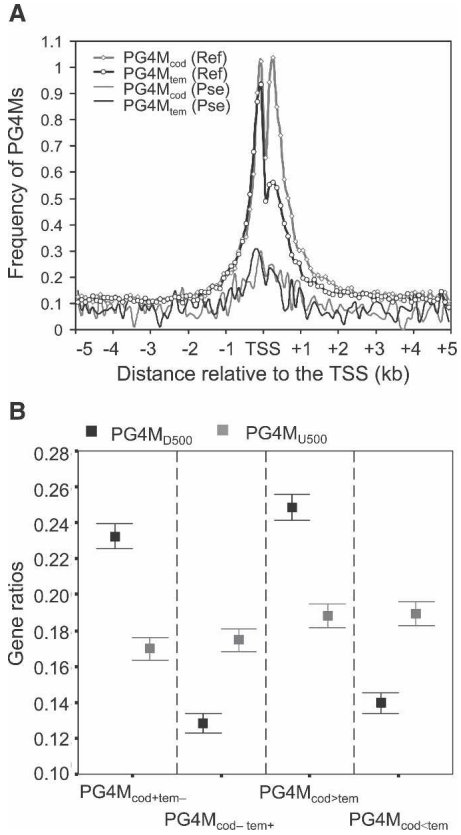

Figure 4. Strand asymmetry of PG4M$_{D500}$. (*A*) Comparison of the frequency between PG4M$_{cod}$ (gray line) and PG4M$_{tem}$ (black line) in the10-kb TSS-flanking region of human RefSeq genes (Ref) (n = 13,276) and pseudogenes (Pse) (n = 824). (*B*) Ratios of genes containing (1) PG4M$_{cod}$ but not PG4M$_{tem}$ (PG4M$_{cod+tem-}$); (2) PG4M$_{tem}$ but not PG4M$_{cod}$ (PG4M$_{cod-tem+}$); (3) more PG4M$_{cod}$ than PG4M$_{tem}$ (PG4M$_{cod>tem}$); and (4) fewer PG4M$_{cod}$ than PG4M$_{tem}$ (PG4M$_{cod<tem}$) in the D500 regions (black squares). The corresponding values for PG4M$_{U500}$ are plotted as a control (gray squares).

cording to their strand location. We found a positive correlation between the frequency of PG4M$_{D500}$ and the mean expression level for both strands (Spearman ρ = 0.538 for PG4M$_{D500cod}$ and gene expression level, $P < 0.001$; Spearman ρ = 0.392 for PG4M$_{D500tem}$ and gene expression level, $P < 0.001$). Additionally, ANOVA results showed that both PG4M$_{D500cod}$-positive and PG4M$_{D500tem}$-positive genes exhibited significantly higher expression levels than the corresponding negative genes (Table 1). To separate there two effects, we performed a multivariate regression analysis with gene expression level as the outcome variable, and the frequencies of PG4M$_{D500cod}$ and PG4M$_{D500tem}$ as the predictor variables. We found that the PG4M$_{D500cod}$ explained 83.5% of the total variance explained by the regression model, while PG4M$_{D500tem}$ explained only the remaining 16.5%. Because the frequencies of PG4M$_{D500cod}$ and PG4M$_{D500tem}$ were correlated (Spearman ρ = 0.282, P = 0.001), to determine whether this could affect our observations, we performed separate regressions on the gene expression levels for each of the variables. The results confirmed the findings of our analysis. These findings thus suggest that PG4M$_{D500}$ on both coding and template strands affect gene expression, with PG4M$_{D500cod}$ being of more importance.

We then investigated whether the asymmetry of PG4M$_{D500}$ affected the level of gene expression. We classified PG4M$_{D500}$-positive genes into three groups (PG4M$_{D500cod>tem}$, n = 3298; PG4M$_{D500cod<tem}$, n = 1854; and PG4M$_{D500cod=tem}$, n = 568). We found that the mean expression level in the PG4M$_{D500cod<tem}$ group (8.850) was lower than that in the other two groups (8.912 and 8.919, respectively), although the differences among the groups were not statistically significant.

PG4M$_{D500}$ is associated with RNA polymerase II occupancy near the TSS

RNA polymerase II (RNAP II), the basal transcription machinery responsible for mRNA synthesis in eukaryotes, is an important factor that is directly associated with gene expression. RNAP II occupancy is an indicator of gene transcription status. Genome-wide studies have demonstrated that the accumulation of RNAP II is largely located at actively transcribed exons (Brodsky et al. 2005), and the TSS-proximal region of a highly expressed gene tends to have a higher RNAP II occupancy (Barski et al. 2007). With the availability of high-resolution profiling of RNAP II binding patterns for human CD4$^+$ T cells (Barski et al. 2007), we compared the average RNAP II occupancy at the 1-kb window around the annotated TSS, among genes containing different numbers of PG4M$_{D500}$ (0, 1, 2, and ≥ 3). ANOVA results showed a significant difference in RNAP II occupancy (log(X+1)-transformed) between PG4M$_{D500}$-positive and PG4M$_{D500}$-negative genes; in addition, the value increased with an increasing number of PG4M$_{D500}$ (Table 2). The median RNAP II occupancy (untransformed data) for PG4M$_{D500}$-positive genes was ~2.1-fold as compared with the corresponding value for PG4M$_{D500}$-negative genes. Since the PG4M$_{D500}$ showed strand asymmetry, we next examined whether this asymmetry influ-

Table 2. Relationship between PG4M$_{D500}$ and RNAP II occupancy at the putative TRR

	Number of genes	RNAP II occupancy (log(X+1) transformed)	*F*-value	*P*-value	Significant in post hoc comparisons
Presence of PG4M$_{D500}$					
0	7556	0.676			
1	3581	0.795			0–01
2	1463	0.822			0–2
≥3	676	0.840	73.694	<0.001	0–≥3
Strand asymmetry of PG4M$_{D500}$					
Cod > Tem	3298	0.830			
Cod = Tem	568	0.812			
Cod < Tem	1854	0.756	11.470	<0.001	Cod < Tem – Cod > Tem[a]

[a]The lack of significance in post hoc comparisons between Cod > Tem and Cod = Tem gene groups might have been due to the insufficient number of genes in the Cod = Tem group.

enced RNAP II occupancy at the TSS-proximal region. Interestingly, we found a higher occupancy for $PG4M_{D500cod>tem}$ genes than for $PG4M_{D500cod=tem}$ or $PG4M_{D500cod<tem}$ genes (Table 2). These results indicate that the pattern of $PG4M_{D500}$ (frequency and strand asymmetry) is associated with an increased RNAP II occupancy near the TSS, supporting a stimulatory role of $PG4M_{D500}$ in transcription.

Discussion

Identification and analysis of functional genomic elements are essential to a better understanding of gene regulation. The ENCyclopedia Of DNA Elements (ENCODE) Project is one such effort, which aims to identify all structural and functional elements encoded in the human genome (The ENCODE Project Consortium 2004, 2007). G4 DNA motifs are widely considered as structural motifs involved in the regulation of transcription, although the underlying molecular mechanism has not been well-characterized. In this study, we found that (1) the presence of potential G4 DNA motifs in the D500 region ($PG4M_{D500}$) increases gene expression level, (2) $PG4M_{D500}$ exhibits strand asymmetry ($PG4M_{cod}$ > $PG4M_{tem}$), which is coupled with the enrichment of PG4Ms in this region, and (3) the presence of $PG4M_{D500}$ and its strand asymmetry are associated with RNAP II occupancy. On the basis of these findings, we propose a model of G4 DNA-based stimulation of transcription (Fig. 5) with the hypothesis that $PG4M_{D500}$ facilitates gene transcription by generating an open DNA structure.

A model of G4 DNA-mediated stimulation of transcription

In the process of transcription, double-stranded DNA is denatured locally by RNAP II, forming a transient structure of ~12–14 bp. This is known as the transcription bubble, which makes the template strand available for nascent RNA synthesis (Zaychikov et al. 1995; Korzheva et al. 2000; Greive and von Hippel 2005). As RNAP II moves along the template strand, the bubble moves with it, and the DNA rewinds to form a duplex behind the bubble. The progression of basal transcription machinery along the template strand will also produce positive and negative supercoiling ahead of and behind the moving complex, respectively (Liu and Wang 1987). Such transcription-derived negative supercoiling can then trigger the unwinding of a double-stranded DNA in the wake of RNAP II. The first scenario in our model is that the stability of double-stranded DNA is a major impediment to the formation of G4 DNA in vivo; hence, the transient separation of duplex DNA as a result of transcription considerably increases the opportunity for the formation of G4 DNA. Once formed, this high-order structure is extremely stable and dissociates slowly (Sen and Gilbert 1990; Mergny et al. 2005). Extensive formation of G4 DNAs on both strands of the D500 region (8830 PG4Ms) can thus impede the renaturation of the duplex, holding the DNA structure open and thereby rendering the template strand available for a higher rate of transcription. A previous in vitro study suggested that the stabilization of duplex DNA by quinone diimine ligands resulted in a significant decrease in the efficiency of transcription by T7 RNA polymerase (Fu et al. 2003). Thus, maintenance of an open DNA structure by G4 DNA may facilitate gene transcription.

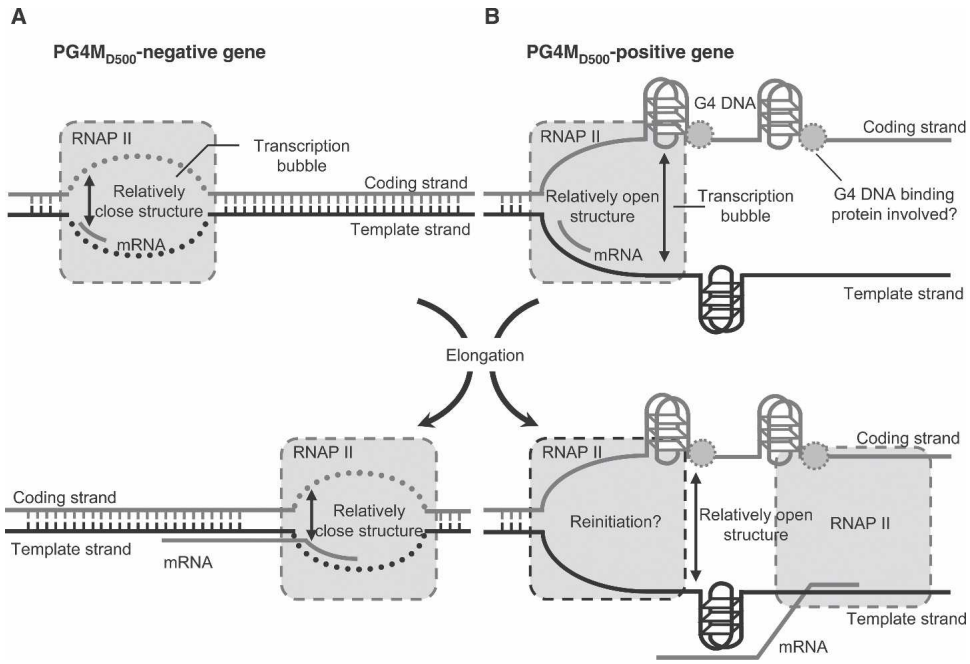

Figure 5. A model of G4 DNA-mediated stimulation of transcription. Double-stranded DNA is denatured locally during transcription, forming a transcription bubble and exposing the template strand for nascent RNA synthesis (*A, top*). As RNAP II machinery (gray dashed box) moves along the template strand, the bubble moves with it, and the DNA rewinds to form the duplex behind the bubble (*A, bottom*). For the $PG4M_{D500}$-positive genes, the transient separation of the duplex DNA during transcription and transcription-derived negative supercoiling considerably increases the opportunities for the formation of G4 DNA. Such high-order structures are extremely stable, and thus, probably block rehybridization with the complementary strand, holding the structure open and thereby enabling a higher rate of transcription (*B*). For the same reason, the presence of high numbers of $PG4M_{D500}$ can help maintain the initial region of the transcript unpaired, which could facilitate the reinitiation of transcription (black dashed box) and also contribute to a higher level of transcription (*B, bottom*). The presence of PG4Ms in the template strand may hinder the progression of RNAP II. However, the intrinsic asymmetric distribution of PG4Ms between strands in the D500 region ($PG4M_{D500cod}$ > $PG4M_{D500tem}$, represented in gray and black, respectively; see also Fig. 4) achieves a balance and minimizes the negative effects.

A high level of gene transcription also depends on a high rate of reinitiation of transcription. In vitro studies have indicated that after the initiation of transcription, a subset of transcription machinery essential for initiation, including general factors, activators, and coactivators, remains bound at the promoter. These dissociated factors form a platform that recruits RNAP II and other factors to initiate a new round of transcription (reinitiation) (Roberts et al. 1995; Zawel et al. 1995; Yudkovsky et al. 2000). Thus, the second scenario in our model is that the open structure resulting from PG4Ms in the initial region of the transcripts may facilitate multiple rounds of transcription.

A barrierless template strand would favor efficient transcription. Although the presence of $PG4M_{D500}$ in either strand would help generate an open DNA structure for transcription by impeding the renaturation of the DNA duplex, these high-order structures, particularly those located on the template strand ($PG4M_{D500tem}$), may also potentially hamper the copying of genetic information onto the mRNA and increase the pause rate, thereby decreasing the progression of the transcription elongation complex. It is thus biologically important to achieve a balance in $PG4M_{D500}$. Consistently, although the presence of $PG4M_{D500tem}$ was found to increase gene expression (Table 1) in this study, genes that contain more $PG4M_{D500cod}$ than $PG4M_{D500tem}$ seem to be expressed at higher levels than those containing more $PG4M_{D500tem}$ than $PG4M_{D500cod}$. According to our data, strand asymmetry of $PG4M_{D500}$ is coupled with enrichment of $PG4M_{D500}$ (Fig. 4A; Supplemental Fig. S2); the number of $PG4M_{D500}$ in the template strand is only ~60% of that in the coding strand, and regression analysis indicated $PG4M_{D500cod}$ having a relatively greater importance than $PG4M_{D500tem}$ in gene expression. Thus, the third scenario in our model is that this asymmetry of $PG4M_{D500}$ may be to minimize the inhibitory effect of $PG4M_{D500}$ on transcription.

The formation of DNA secondary structures during transcription has been well-documented previously. The unwinding of the double-stranded DNA structures during transcription provides the conditions required for the formation of DNA secondary structures cotranscriptionally, for example, the stem–loop structures formed in single-stranded DNA (ssDNA) (Dayn et al. 1992; Wright 2004; Drolet 2006) and the R-loop formed in the G-rich immunoglobulin class switch regions (Yu et al. 2003; Chaudhuri and Alt 2004), where the RNA strand is base paired with the template strand, leaving the G-rich coding strand displaced. Although the extent and prevalence of the existence of G4 DNA in vivo need further characterization, the formation of a G4 DNA structure during transcription has been documented (Duquette et al. 2004). It is also possible that some regulatory factors specifically bind to G4 DNA and stabilize it, maintaining the DNA structure in an open conformation. An example of G4 DNA-mediated stimulation of transcription is rRNA synthesis. The coding strand of the rDNA gene is rich in guanine and can form G4 DNA. Results of a previous study suggested that by interacting with nucleolin, the G4 DNA formed in the coding strand may contribute to a high rate of rRNA transcription by making the template strand available for multiple rounds of transcription (Hanakahi et al. 1999; Maizels 2006). At the same time, as there are some transcription factors such as BGP1 (also known as *VEZF1*) (Lewis et al. 1988; Clark et al. 1990; Dexheimer et al. 2006) and MAZ (also known as Pur-1) (Kennedy and Rutter 1992; Catasti et al. 1996; Lew et al. 2000; Dexheimer et al. 2006) that can activate gene expression through recognition and binding of G4 DNA motifs formed in the regulatory region, some ubiquitous regulatory proteins may also be involved in and contribute to the mechanisms of G4 DNA-mediated transcriptional activation.

The RNAP II occupancy analysis in the 1-kb window around the TSS (–500 to +500) provides support for our model. If $PG4M_{D500}$ and its strand asymmetry contributed to a high rate of transcription and to multiple rounds of transcription by generating an open DNA structure, we would expect an increase in RNAP II occupancy for $PG4M_{D500}$-positive genes and for genes with more $PG4M_{D500}$ in the coding strand. As expected, $PG4M_{D500}$-positive genes and genes with such asymmetry of $PG4M_{D500}$ showed significantly higher RNAP II occupancy in the TSS-proximal region than $PG4M_{D500}$-negative genes and genes without such asymmetry (Table 2).

Together, in this model we hypothesized that the presence of PG4Ms (both $PG4M_{cod}$ and $PG4M_{tem}$) in the D500 region contributes to a higher rate of transcription and facilitates multiple rounds of transcription by generating an open DNA structure through the formation of G4 DNA. Additionally, the strand asymmetry of PG4Ms ($PG4M_{tem} < PG4M_{cod}$) in the D500 region reduces the potential inhibitory effect of the G4 DNA structure on the progression of RNAP II.

Potential roles of PG4Ms in regions outside the immediate downstream region

In this study, the putative TRR (TSS ± 500) was chosen for analysis because PG4Ms are significantly enriched in this region. Our model suggests a stimulatory role in transcription for PG4Ms, in the initial region of the transcripts (TSS to +500, D500). A further question is whether this model can be extended to entire transcripts. We repeated our analysis for the rest of the transcripts (+500 to the transcription termination site). Interestingly, a correlation between the frequency of PG4Ms and the gene expression level was observed in this region, even after controlling for the effect of the D500 region (Spearman partial $\rho = 0.437$, $P < 0.001$). This indicated that PG4Ms may facilitate gene transcription along the entire transcription region, consistent with scenario one of our hypothesis. However, because the frequency of PG4Ms and their strand asymmetry were greatly decreased in the rest of the transcription region (data not shown), we suggest that PG4Ms in the remaining region are not as important as those in the initial region of the transcript.

Although $PG4M_{U500}$ showed no association with expression levels in this study, a role in the regulation of gene transcription cannot be ruled out. Compared with pseudogenes, the frequency of PG4Ms is significantly great in the U500 region of the intact RefSeq genes (1.35 vs. 0.42). Moreover, several experimental investigations have demonstrated that the G4 DNA formed in promoters play an important role in the regulation of transcription (Lewis et al. 1988; Clark et al. 1990; Lew et al. 2000; Siddiqui-Jain et al. 2002; Cogoi and Xodo 2006; Dexheimer et al. 2006). $PG4M_{U500}$ may influence the assembly of the transcription initiation complex by recruiting or blocking the binding of transcription factors, thereby regulating the initiation of transcription.

Concluding remarks

In this study, we have performed a comprehensive analysis of the relationship between potential G4 DNA motifs in the putative transcriptional regulatory region and gene expression level. Our findings suggest that the presence of PG4Ms and their intrinsic asymmetry in the 500-bp downstream region with respect to the TSS are associated with the regulation of gene expression. At

present, the exact role of G4 DNA motifs in transcription is not well-characterized, and the model proposed here needs to be validated by further experimental studies. Nonetheless, this is the first genome-wide analysis detailing the potential role of G4 DNA in gene transcription. Our results provide information that is essential for a global understanding of the regulatory role of G4 DNA in the human genome.

Methods

Data sources

DNA sequences

Genomic sequences for both the TSS-flanking region and the transcription region were extracted from the UCSC genome browser (hg 18) (Karolchik et al. 2004). The IDs of human pseudogenes were obtained from Ensembl (Kasprzyk et al. 2004). For genes with two or more transcripts, one transcript was selected at random for analysis.

Gene expression data

Gene expression data for 79 human tissues/cells were retrieved from the Gene Atlas V2 data set standardized with the MAS5 algorithm (Su et al. 2004). Expression values from the probe sets corresponding to the same gene were averaged. The expression values were then log2 transformed to approximate a normal fit.

Gene family

Information on gene families was extracted from the HUGO Gene Nomenclature Committee (Eyre et al. 2006). Only gene families that contained 10 or more members were selected for analysis.

Gene functional annotation

The functional annotations for human genes were retrieved from the PANTHER classification system (Thomas et al. 2003).

RNAP II occupancy

The high-resolution genome-wide RNAP II occupancy in the human genome was obtained from the work of Artem Barski et al. (Barski et al. 2007). More information on this data is described elsewhere. The average RNAP II occupancy (number of tags per kilobase) at the 1-kb window around the TSS was calculated by mapping it (hg 18) to the tags in 5-bp windows. The values were finally log(x+1) transformed following the indications of the Box-Cox method.

Intramolecular G4 DNA identification

The Quadparser program (Huppert and Balasubramanian 2005) was adopted for intramolecular G4 DNA identification, with default parameters. Typically, to be recognized as a putative G4 DNA-forming site, a DNA sequence must have four or more G-runs, and each G-run must contain three or more continuous guanine bases, leading to a typical folding rule: $G_{\geq 3}N_{1-7}G_{\geq 3}N_{1-7}G_{\geq 3}N_{1-7}G_{\geq 3}$, where N refers to any base and serves as the loop that connects the G-runs. As the sequences are provided in the single-stranded form, the default parameters of Quadparser identify PG4Ms in both DNA strands. To distinguish PG4Ms in the coding and template strands, the modified parameters "G 3 4 1 7" and "C 3 4 1 7" were used. The PG4Ms mentioned in this study refer to distinct PG4M-forming sites, each of which could form diverse topologies of G4 DNA.

Statistics

We used ANOVA to test for statistical differences in the mean gene expression level and RNAP II occupancy among the PG4M groups, followed by the Tukey post hoc test if multiple comparisons were needed. Because gene expression levels vary from tissue to tissue, we also performed multivariate analysis of variance, which took into account expression variation among tissues/cells. The multivariate test statistic Pillai's Trace was used because it is robust with respect to both departures from homogeneity and multivariate normality. To disentangle the effects of $PG4M_{D500cod}$ and $PG4M_{D500tem}$ on gene expression, we used multivariate regression, with a forward stepwise selection procedure. To avoid the problem that many of the individual genes had a frequency of PG4M of 0, we sorted 13,276 genes according to their expression level and computed the average expression level and average frequency of PG4Ms for every 100-gene window. Spearman rank correlation was then used to test the trend of the relationship between these two variables. The same trend was obtained using different window sizes, so we only presented the results obtained using the 100-gene window. It should be noted that this binned method might decrease the variance of the original data, so it could not be used to quantify their relationship, but to discover the trend of the relationship.

Acknowledgments

This work was supported by the National Major Basic Research Program of China (973 program, No. 2006CB102100), the National High Technology Research and Development Program of China (863 Program, No. 2006AA10A120), and the National Natural Science Foundation of China. We express our sincere gratitude to four anonymous referees for their constructive comments and suggestions.

References

Ambrus, A., Chen, D., Dai, J., Jones, R.A., and Yang, D. 2005. Solution structure of the biologically relevant G-quadruplex element in the human c-MYC promoter. Implications for G-quadruplex stabilization. *Biochemistry* **44:** 2048–2058.

Barski, A., Cuddapah, S., Cui, K., Roh, T.Y., Schones, D.E., Wang, Z., Wei, G., Chepelev, I., and Zhao, K. 2007. High-resolution profiling of histone methylations in the human genome. *Cell* **129:** 823–837.

Brodsky, A.S., Meyer, C.A., Swinburne, I.A., Hall, G., Keenan, B.J., Liu, X.S., Fox, E.A., and Silver, P.A. 2005. Genomic mapping of RNA polymerase II reveals sites of co-transcriptional regulation in human cells. *Genome Biol.* **6:** R64. doi: 10.1186/gb-2005-6-8-r64.

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. doi: 10.1093/nar/gkl655.

Catasti, P., Chen, X., Moyzis, R.K., Bradbury, E.M., and Gupta, G. 1996. Structure-function correlations of the insulin-linked polymorphic region. *J. Mol. Biol.* **264:** 534–545.

Chaudhuri, J. and Alt, F.W. 2004. Class-switch recombination: Interplay of transcription, DNA deamination and DNA repair. *Nat. Rev. Immunol.* **4:** 541–552.

Clark, S.P., Lewis, C.D., and Felsenfeld, G. 1990. Properties of BGP1, a poly(dG)-binding protein from chicken erythrocytes. *Nucleic Acids Res.* **18:** 5119–5126. doi: 10.1093/nar/18.17.5119.

Cogoi, S. and Xodo, L.E. 2006. G-quadruplex formation within the promoter of the KRAS proto-oncogene and its effect on transcription. *Nucleic Acids Res.* **34:** 2536–2549. doi: 10.1093/nar/gkl286.

Dai, J., Chen, D., Jones, R.A., Hurley, L.H., and Yang, D. 2006. NMR solution structure of the major G-quadruplex structure formed in the human BCL2 promoter region. *Nucleic Acids Res.* **34:** 5133–5144. doi: 10.1093/nar/gkl610.

Dayn, A., Malkhosyan, S., and Mirkin, S.M. 1992. Transcriptionally driven cruciform formation in vivo. *Nucleic Acids Res.* **20:** 5991–5997. doi: 10.1093/nar/20.22.5991.

De Armond, R., Wood, S., Sun, D., Hurley, L.H., and Ebbinghaus, S.W.

2005. Evidence for the presence of a guanine quadruplex forming region within a polypurine tract of the hypoxia inducible factor 1alpha promoter. *Biochemistry* **44:** 16341–16350.
Dexheimer, T.S., Fry, M., and Hurley, L.H. 2006. DNA quadruplexes and gene regulation. In *Quadruplex nucleic acids* (eds. S. Neidle and S. Balasubramanian), pp. 180–207. RSC Publishing, Cambridge, UK
Drolet, M. 2006. Growth inhibition mediated by excess negative supercoiling: The interplay between transcription elongation, R-loop formation and DNA topology. *Mol. Microbiol.* **59:** 723–730.
Du, Z., Kong, P., Gao, Y., and Li, N. 2007. Enrichment of G4 DNA motif in transcriptional regulatory region of chicken genome. *Biochem. Biophys. Res. Commun.* **354:** 1067–1070.
Duquette, M.L., Handa, P., Vincent, J.A., Taylor, A.F., and Maizels, N. 2004. Intracellular transcription of G-rich DNAs induces formation of G-loops, novel structures containing G4 DNA. *Genes & Dev.* **18:** 1618–1629.
The ENCODE Project Consortium. 2004. The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science* **306:** 636–640.
The ENCODE Project Consortium. 2007. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* **447:** 799–816.
Eyre, T.A., Ducluzeau, F., Sneddon, T.P., Povey, S., Bruford, E.A., and Lush, M.J. 2006. The HUGO Gene Nomenclature Database, 2006 updates. *Nucleic Acids Res.* **34:** D319–D321. doi: 10.1093/nar/gkj147.
Fanger, N.A., Borges, L., and Cosman, D. 1999. The leukocyte immunoglobulin-like receptors (LIRs): A new family of immune regulators. *J. Leukoc. Biol.* **66:** 231–236.
Fernando, H., Reszka, A.P., Huppert, J., Ladame, S., Rankin, S., Venkitaraman, A.R., Neidle, S., and Balasubramanian, S. 2006. A conserved quadruplex motif located in a transcription activation site of the human c-kit oncogene. *Biochemistry* **45:** 7854–7860.
Fry, M. 2007. Tetraplex DNA and its interacting proteins. *Front. Biosci.* **12:** 4336–4351.
Fu, P.K., Bradley, P.M., and Turro, C. 2003. Stabilization of duplex DNA structure and suppression of transcription in vitro by bis(quinone diimine) complexes of rhodium(III) and ruthenium(II). *Inorg. Chem.* **42:** 878–884.
Gellert, M., Lipsett, M.N., and Davies, D.R. 1962. Helix formation by guanylic acid. *Proc. Natl. Acad. Sci.* **48:** 2013–2018.
Grand, C.L., Han, H., Munoz, R.M., Weitman, S., Von Hoff, D.D., Hurley, L.H., and Bearss, D.J. 2002. The cationic porphyrin TMPyP4 down-regulates c-MYC and human telomerase reverse transcriptase expression and inhibits tumor growth in vivo. *Mol. Cancer Ther.* **1:** 565–573.
Greive, S.J. and von Hippel, P.H. 2005. Thinking quantitatively about transcriptional regulation. *Nat. Rev. Mol. Cell Biol.* **6:** 221–232.
Guschlbauer, W., Chantot, J.F., and Thiele, D. 1990. Four-stranded nucleic acid structures 25 years later: From guanosine gels to telomer DNA. *J. Biomol. Struct. Dyn.* **8:** 491–511.
Han, H. and Hurley, L.H. 2000. G-quadruplex DNA: A potential target for anti-cancer drug design. *Trends Pharmacol. Sci.* **21:** 136–142.
Hanakahi, L.A., Sun, H., and Maizels, N. 1999. High affinity interactions of nucleolin with G-G-paired rDNA. *J. Biol. Chem.* **274:** 15908–15912.
Huppert, J.L. and Balasubramanian, S. 2005. Prevalence of quadruplexes in the human genome. *Nucleic Acids Res.* **33:** 2908–2916. doi: 10.1093/nar/gki609.
Huppert, J.L. and Balasubramanian, S. 2007. G-quadruplexes in promoters throughout the human genome. *Nucleic Acids Res.* **35:** 406–413. doi: 10.1093/nar/gkl1057.
Karolchik, D., Hinrichs, A.S., Furey, T.S., Roskin, K.M., Sugnet, C.W., Haussler, D., and Kent, W.J. 2004. The UCSC Table Browser data retrieval tool. *Nucleic Acids Res.* **32:** D493–D496. doi: 10.1093/nar/gkh103.
Kasprzyk, A., Keefe, D., Smedley, D., London, D., Spooner, W., Melsopp, C., Hammond, M., Rocca-Serra, P., Cox, T., and Birney, E. 2004. EnsMart: A generic system for fast and flexible access to biological data. *Genome Res.* **14:** 160–169.
Keniry, M.A. 2000. Quadruplex structures in nucleic acids. *Biopolymers* **56:** 123–146.
Kennedy, G.C. and Rutter, W.J. 1992. Pur-1, a zinc-finger protein that binds to purine-rich sequences, transactivates an insulin promoter in heterologous cells. *Proc. Natl. Acad. Sci.* **89:** 11498–11502.
Korzheva, N., Mustaev, A., Kozlov, M., Malhotra, A., Nikiforov, V., Goldfarb, A., and Darst, S.A. 2000. A structural model of transcription elongation. *Science* **289:** 619–625.
Lew, A., Rutter, W.J., and Kennedy, G.C. 2000. Unusual DNA structure of the diabetes susceptibility locus IDDM2 and its effect on transcription by the insulin promoter factor Pur-1/MAZ. *Proc. Natl. Acad. Sci.* **97:** 12508–12512.
Lewis, C.D., Clark, S.P., Felsenfeld, G., and Gould, H. 1988. An erythrocyte-specific protein that binds to the poly(dG) region of the chicken beta-globin gene promoter. *Genes & Dev.* **2:** 863–873.
Liu, L.F. and Wang, J.C. 1987. Supercoiling of the DNA template during transcription. *Proc. Natl. Acad. Sci.* **84:** 7024–7027.
Maizels, N. 2006. Dynamic roles for G4 DNA in the biology of eukaryotic cells. *Nat. Struct. Mol. Biol.* **13:** 1055–1059.
Mergny, J.L., De Cian, A., Ghelab, A., Sacca, B., and Lacroix, L. 2005. Kinetics of tetramolecular quadruplexes. *Nucleic Acids Res.* **33:** 81–94. doi: 10.1093/nar/gki148.
Rankin, S., Reszka, A.P., Huppert, J., Zloh, M., Parkinson, G.N., Todd, A.K., Ladame, S., Balasubramanian, S., and Neidle, S. 2005. Putative DNA quadruplex formation within the human c-kit oncogene. *J. Am. Chem. Soc.* **127:** 10584–10589.
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.
Roberts, S.G., Choy, B., Walker, S.S., Lin, Y.S., and Green, M.R. 1995. A role for activator-mediated TFIIB recruitment in diverse aspects of transcriptional regulation. *Curr. Biol.* **5:** 508–516.
Seenisamy, J., Rezler, E.M., Powell, T.J., Tye, D., Gokhale, V., Joshi, C.S., Siddiqui-Jain, A., and Hurley, L.H. 2004. The dynamic character of the G-quadruplex element in the c-MYC promoter and modification by TMPyP4. *J. Am. Chem. Soc.* **126:** 8702–8709.
Sen, D. and Gilbert, W. 1990. A sodium-potassium switch in the formation of four-stranded G4-DNA. *Nature* **344:** 410–414.
Shafer, R.H. and Smirnov, I. 2000. Biological aspects of DNA/RNA quadruplexes. *Biopolymers* **56:** 209–227.
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.* **99:** 11593–11598.
Simonsson, T. 2001. G-quadruplex DNA structures–variations on a theme. *Biol. Chem.* **382:** 621–628.
Su, A.I., Wiltshire, T., Batalov, S., Lapp, H., Ching, K.A., Block, D., Zhang, J., Soden, R., Hayakawa, M., Kreiman, G., et al. 2004. A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc. Natl. Acad. Sci.* **101:** 6062–6067.
Sun, D., Guo, K., Rusche, J.J., and Hurley, L.H. 2005. Facilitation of a structural transition in the polypurine/polypyrimidine tract within the proximal promoter region of the human VEGF gene by the presence of potassium and G-quadruplex-interactive agents. *Nucleic Acids Res.* **33:** 6070–6080. doi: 10.1093/nar/gki917.
Thomas, P.D., Campbell, M.J., Kejariwal, A., Mi, H., Karlak, B., Daverman, R., Diemer, K., Muruganujan, A., and Narechania, A. 2003. PANTHER: A library of protein families and subfamilies indexed by function. *Genome Res.* **13:** 2129–2141.
Todd, A.K., Johnston, M., and Neidle, S. 2005. Highly prevalent putative quadruplex sequence motifs in human DNA. *Nucleic Acids Res.* **33:** 2901–2907. doi: 10.1093/nar/gki553.
Wright, B.E. 2004. Stress-directed adaptive mutations and evolution. *Mol. Microbiol.* **52:** 643–650.
Xu, Y. and Sugiyama, H. 2006. Formation of the G-quadruplex and i-motif structures in retinoblastoma susceptibility genes (Rb). *Nucleic Acids Res.* **34:** 949–954. doi: 10.1093/nar/gkj485.
Yafe, A., Etzioni, S., Weisman-Shomer, P., and Fry, M. 2005. Formation and properties of hairpin and tetraplex structures of guanine-rich regulatory sequences of muscle-specific genes. *Nucleic Acids Res.* **33:** 2887–2900. doi: 10.1093/nar/gki606.
Yu, K., Chedin, F., Hsieh, C.L., Wilson, T.E., and Lieber, M.R. 2003. R-loops at immunoglobulin class switch regions in the chromosomes of stimulated B cells. *Nat. Immunol.* **4:** 442–451.
Yudkovsky, N., Ranish, J.A., and Hahn, S. 2000. A transcription reinitiation intermediate that is stabilized by activator. *Nature* **408:** 225–229.
Zawel, L., Kumar, K.P., and Reinberg, D. 1995. Recycling of the general transcription factors during RNA polymerase II transcription. *Genes & Dev.* **9:** 1479–1490.
Zaychikov, E., Denissova, L., and Heumann, H. 1995. Translocation of the *Escherichia coli* transcription complex observed in the registers 11 to 20: "Jumping" of RNA polymerase and asymmetric expansion and contraction of the "transcription bubble". *Proc. Natl. Acad. Sci.* **92:** 1739–1743.
Zhao, Y., Du, Z., and Li, N. 2007. Extensive selection for the enrichment of G4 DNA motifs in transcriptional regulatory regions of warm blooded animals. *FEBS Lett.* **581:** 1951–1956.

Received July 12, 2007; accepted in revised form October 24, 2007.