FEBS Letters 581 (2007) 1951–1956

# Extensive selection for the enrichment of G4 DNA motifs in transcriptional regulatory regions of warm blooded animals

Yiqiang Zhao[1], Zhuo Du[1], Ning Li*

*State Key Laboratory for Agrobiotechnology, China Agricultural University, Beijing, 10094, China*

Received 31 January 2007; revised 2 April 2007; accepted 3 April 2007

Available online 18 April 2007

Edited by Takashi Gojobori

**Abstract** **A comprehensive analysis of potential G4 DNA motifs (G4Ms) in genomic regions flanking transcription start sites (TSS) was performed across 13 animal species. We found that G4Ms are significantly enriched in the transcriptional regulatory regions (TRRs) of warm-blooded animals. Further analysis of human genes in different temporal groups reveals that the enrichment is not specific to genes found only in warm-blooded species but instead exist in a wide range of genes. Our findings therefore suggest that the high prevalence of G4Ms in TRRs is extensively selected in warm-blooded animals, supporting the hypothesis that G4Ms are involved in the regulation of gene transcription.**


*Keywords:* G-quadruplex; G4 DNA; Transcription; Transcription start site; Transcriptional regulatory region; Warm-blooded animal

## 1. Introduction

The G-quadruplex, or G4 DNA which is formed in guanine-rich sequences, is a stable, four-stranded DNA structure alternative to the conventional double-stranded conformation [1–4]. To form G4 DNA, four guanine residues are Hoogsteen hydrogen-bonded in a square planar array called G-quartets; three or more quartets are then stacked on one another and connected by three intervening loops [1–4].

Although G4 DNA has been recognized for more than 40 years, its biological function is still not well characterized [3,5,6]. Several recent studies, however, have provided novel insight into the potential function of G4 DNA motifs (G4Ms) in regulating gene expression [2,6–12]. For instance, a chair-formed intramolecular G4 DNA identified upstream of the P1 promoter of the human c-MYC gene was shown to considerably repress c-MYC transcription [7,11]. Conversely, G4 DNA formed in the polymorphic G-rich mini-satellite located upstream of the human insulin gene has been reported to be important for its transcriptional activation [12]. Additionally, a number of G4M-forming sequences have been subsequently identified in the promoters of several important cancer-related genes, including c-Kit, Bcl-2, VEGF, KRAS and RB [13–19]. This combined evidence suggests that G4Ms may be common elements involved in transcriptional regulation. Furthermore, genomic analysis has revealed that potential G4M-forming sequences are highly prevalent in the human genome [20,21] and more importantly, are significantly enriched in promoters of both the human and chicken genomes [8,10]. All of these findings support the current hypothesis that G4Ms may be a novel type of regulatory element that contributes to gene expression regulation through a structural-mediated mechanism.

Although the potential role for G4Ms in gene regulation has been previously investigated, relatively few studies have examined this topic in depth. Therefore, in this study, we chose to analyze and compare, in more detail, the distribution of G4Ms in genomic regions flanking the transcription start site (TSS) of 13 animal species. Our results suggest an extensive selection for G4M-enrichment in transcriptional regulatory regions (TRRs) of warm-blooded animals. These findings will be helpful in furthering the understanding of the regulatory function of G4Ms in gene expression.

## 2. Materials and methods

### 2.1. Data sets

Ten kilobases of genomic sequences flanking the TSS (±5 kb) of 13 species were retrieved from UCSC genome browser[22] including human (*Homo sapiens*), chimpanzee (*Pan troglodytes*), rat (*Rattus norvegicus*), mouse (*Mus musculus*), dog (*Canis familiaris*), cattle (*Bos Taurus*) , chicken (*Gallus gallus*), tropical clawed frog (*Xenopus tropicalis*), fugu (*Takifugu rubripes*), water fresh pufferfish (*Tetraodon nigroviridis*), Zebrafish (*Danio rerio Tuebingen*), fly (*Drosophila melanogaster*) and nematode (*Caenorhabditis elegans*). General information for these datasets were listed in Table 1.

### 2.2. Identification of putative G4 DNA motifs

The Quadparser program developed by Julian L. Huppert and coworkers [21] was applied in this study to identify putative G4 DNA with default parameters. To be recognized as a potential G4M-forming site, a sequence had to comply with the following folding rule: $G_{\geqslant 3}N_{1-7}G_{\geqslant 3}N_{1-7}G_{\geqslant 3}N_{1-7}G_{\geqslant 3}$, where N refers to any base. Because genomic DNA is provided as a single-strand, both G- and C-patterns were searched in order to identify potential G4 DNA on both strands. The Quadparser program only assigns one count to a sequence, regardless of how many potential G4 DNA motifs could form in a given G4M-forming sequence; therefore, the G4 DNA motifs mentioned in this study represent distinct sites, each with the potential to form G4 DNA. Detailed instructions for using this program are fully described elsewhere [21].

*Corresponding author. Fax: +86 10 62733904.
*E-mail address:* ninglbau@public3.bta.net.cn (N. Li).

[1]These authors contributed equally to this work.

*Abbreviations:* G4M, G4 DNA motif; TSS, transcription start site; TRR, transcriptional regulatory region

0014-5793/$32.00 
doi:10.1016/j.febslet.2007.04.017

Table 1
General information of G4 DNA motif in 13 analyzed animal genomes

| Species | Gene no. | G4M no. | Frequency | Up100 frequency | G4M gene | Ratio (%) | Source |
|---|---|---|---|---|---|---|---|
| Human | 25099 | 99369 | 0.40 | 1.84 | 14347 | 57.2 | Refseq genes |
| Chimpanzee | 25410 | 85299 | 0.34 | 1.38 | 13481 | 53.1 | Refseq genes |
| Rat | 10105 | 24367 | 0.24 | 0.80 | 3764 | 37.2 | Refseq genes |
| Mouse | 19862 | 59117 | 0.30 | 1.28 | 9196 | 46.3 | Refseq genes |
| Dog | 770 | 4135 | 0.54 | 1.17 | 405 | 52.6 | Refseq genes |
| Cattle | 8243 | 26376 | 0.32 | 1.18 | 3661 | 44.4 | Refseq genes |
| Chicken | 4026 | 13699 | 0.34 | 1.52 | 2274 | 56.5 | Refseq genes |
| Tropical clawed frog | 6142 | 10293 | 0.17 | 0.30 | 1150 | 18.7 | Refseq genes |
| Fugu | 38510 | 44907 | 0.12 | 0.16 | 6986 | 18.1 | Ensembl genes |
| Water fresh pufferfish | 27918 | 38398 | 0.14 | 0.16 | 3466 | 12.4 | Genoscope GAZE |
| Zebrafish | 12038 | 3352 | 0.03 | 0.04 | 251 | 2.1 | Refseq genes |
| Fruit fly | 21047 | 10506 | 0.05 | 0.02 | 716 | 3.4 | Refseq genes |
| Nematode | 23527 | 4379 | 0.03 | 0.03 | 439 | 1.9 | Refseq genes |

Gene no., the total genes be analyzed; G4M no., the number of G4M identified in 10 kb regions (±5 kb) flanking the TSS; Frequency, the frequency (number of G4M per kb) of G4M; Up100 frequency, the frequency of G4M in 100 bp upstream the TSS; G4M gene, the number of gene contained at least one G4M in the transcription regulatory region (TRR, −500 to +500 bp); Ratio, the percentage of G4M gene; Source, the source of the sequence data.

### 2.3. Classification of human genes

Ortholog sets between human and other species were identified using reciprocal BLAST [23], with an *E*-value threshold set to $E-25$. We classified human genes into temporal groups using known phylogeny between human and fly, fugu, frog, chicken, dog and mouse, and then examined the presence of the genes in each of these groups. Thus, if a gene was present in the human–fly ortholog set, it would be classified into the oldest temporal group. Similarly, if a second gene was located in the human–fugu ortholog set but not in the human–fly ortholog set, the assumption was made that it was introduced at this stage in the phylogenetic chain and was therefore placed in the second temporal group. Nucleotide sequences (cDNA) used to construct ortholog sets were collected from three sources: the NCBI Reference Sequence Database for human, mouse and fly; the Gene Index Database for chicken and frog; and the Ensembl Database for chimpanzee, dog and fugu.

## 3. Results and discussion

### 3.1. High frequency of G4 DNA motifs in TSS-flanking regions of warm-blooded animals

We analyzed the frequency of potential G4Ms in TSS-flanking regions (±5 kb) of 13 sequenced animal genomes, including five mammals, one avian, one amphibian, three fish, one insect and one nematode. As listed in Table 1, the average frequency of G4Ms (number of G4Ms per kilobase of genomic DNA) in TSS-flanking regions ranged from 0.24 to 0.54 in warm-blooded animals (*Note*: the highest value observed in dog might be somewhat biased due to insufficient gene number). In contrast, the frequency was significantly lower in cold-blooded animals than that in warm-blooded animals, from 0.03 to 0.17 (Mann–Whitney test, $P < 0.0001$).

One consideration for such a discrepancy is that the high frequency of G4Ms in warm-blooded animals might simply be due to a higher elevation in GC-content. From our data, warm-blooded animals were significantly richer in GC-regions than cold-blooded ones (Mann–Whitney test, $P < 0.0001$) and a positive correlation between GC-content and frequency of G4Ms exists in all warm-blooded animals tested (data not shown). However, these positive correlations disappeared in most of the cold-blooded animals. More importantly, differences in the frequency of G4Ms between warm- and cold-blooded animals were still significant ($F = 30.847$, $P < 0.0001$) when applying analysis of variance, with the GC-content as covariate.

### 3.2. G4 DNA motifs are enriched in transcriptional regulatory regions of warm-blooded animals

Recent studies have reported the enrichment of G4Ms in the promoter regions of human and chicken genomes [8,13]. Fig. 1 (black lines) exhibits the presence of G4M-enrichment in the transcriptional regulatory regions (TRRs; defined as 1 kb genomic regions flanking the TSS, −500 to +500) of warm-blooded animals. To confirm this, we compared the frequency of G4Ms along the 10 kb TSS-flanking regions by dividing it into five parts (−5000 to −2001; −2000 to −501; −500 to +499; +500 to +1999; and +2000 to +5000). In warm-blooded animals, a significant difference in G4M frequency was detected among these sequence sections (Kruskal–Wallis test, $H = 292.34$; df = 4; $P < 0.0001$) with the highest value located in the TRR region (−500 to +499) (Mann–Whitney test, $P < 0.0001$; Bonferroni-corrected). In particular, the frequency was more remarkable when examining a narrower core promoter region (−100 to the TSS) (see Table 1).

Many seemingly G4M-forming sequences do not actually fold into stable G4 DNA [19,24]; therefore, the frequency of potential stable G4Ms is more biologically relevant. At present, it is difficult to predict which potential topology of G4 DNA formed in a given sequence is more stable. However, it has been reported that (i) the presence of single-nucleotide loops could increase the stability of the G4 DNA; and (ii) single-nucleotide loops can adapt double-chain reversals which drive the G4 DNA to form parallel folds [14,24,25]. We therefore calculated the frequency of stable G4Ms by restricting the loop size to one in at least one of three connecting loops in the warm-blooded animals. As shown in Fig. 1 (grey lines), stable G4Ms also followed an enriched pattern in the TRRs.

Considering that TRRs generally tend to be GC-rich, we next performed a randomization procedure to test whether the high frequency of G4Ms obtained from TRRs could be directly explained by a GC-rich environment. 1000 randomized data sets were created, where gene length and base frequency were held constant but base position was randomly permuted. We calculated the frequency of G4Ms for each randomized set and, if n was the number of random data sets for which the G4M frequency was equal to or greater than the observed G4M frequency in the real TRR sequences, we estimated $P = (n + 1)/1001$. Results show that the frequency of G4Ms

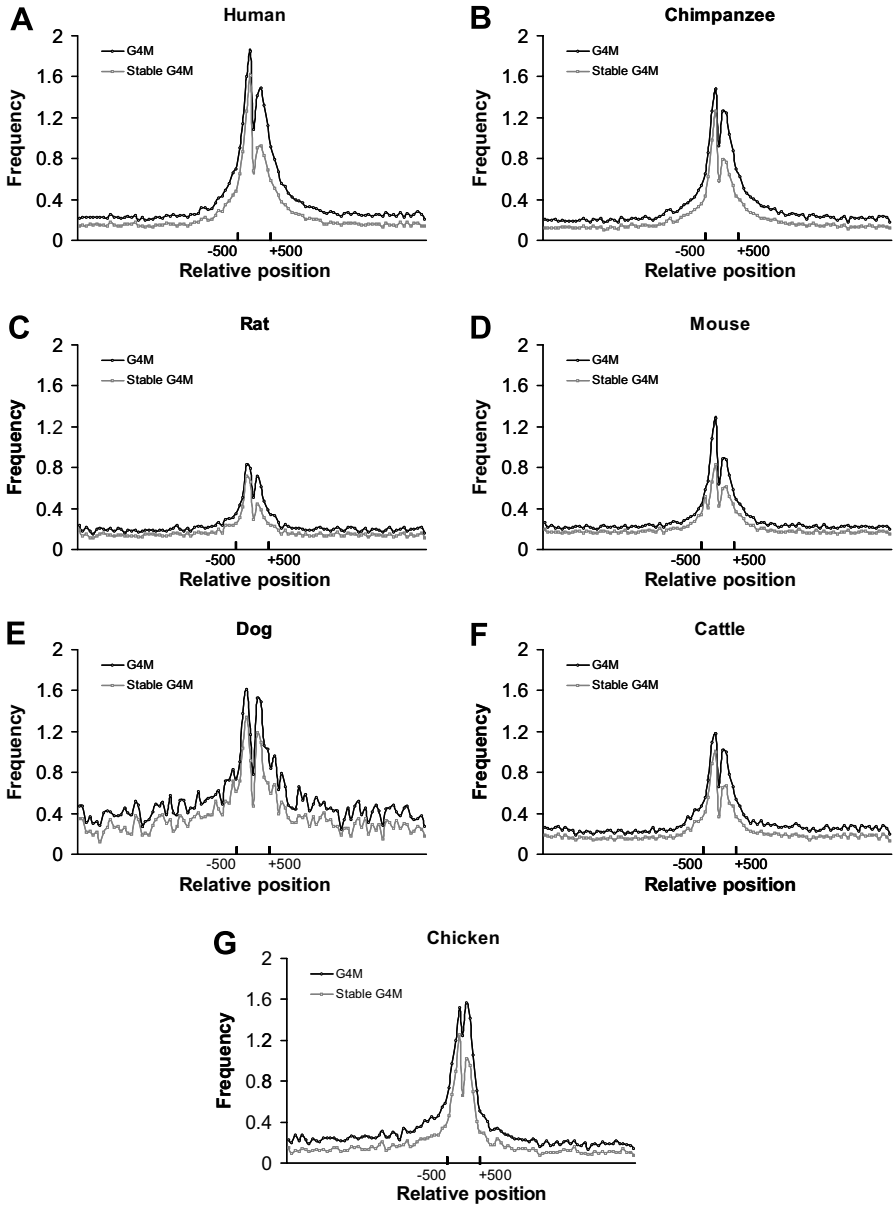


Fig. 1. Distribution of G4Ms in TSS-flanking regions of warm-blooded animals. G4 DNA motifs are not evenly distributed along the TSS-flanking regions but instead are clustered in the TRRs of warm-blooded animals (TSS ± 500). The black lines represent the frequency of overall potential G4Ms in 100 bp windows, whereas the grey lines correspond to stable G4Ms. The position of the TRR is marked.

in TRRs exceeded that in the randomized sets by about 5 times in all warm-blooded animals ($P < 0.001$; data not shown). Thus, it appears that an additional factor must be present that contributes to the observed enrichment of G4Ms in TRRs, in spite of high GC-content.

We also counted the number of genes that had at least one G4 DNA in the TRR (G4M gene). As shown in Table 1, warm-blooded animals had significantly more G4M genes compared to cold-blooded animals ($\chi^2 = 44753.98$; df = 1; $P < 0.0001$). The ratio of G4M genes ranged from 37% to 57% in warm-blooded animals and 1.9% to 18.7% in cold-blooded animals (Table 1). Thus, it is less likely that the higher G4M frequency observed in the TRR region in warm-blooded animals is contributed by a limited set of genes but rather as a consequence of large-scale adoption.

Additionally, Fig. 1 clearly shows a double-peak distribution of G4Ms in warm-blooded animals: the first peak appears in the region −100 to the TSS; the frequency slightly decreases along a 150 bp region downstream of the TSS. The second peak appears in the region +150 to +300. Neither the G4M-enriched TRRs, nor the double-peak feature, were found in cold-blooded animals (Fig. 2).

Taken together, these findings suggest that overall G4Ms, along with the stable forms, are widely adopted in the TRRs of warm-blooded animals. This supports the hypothesis that G4 DNA might be a novel type of element associated with transcriptional regulation. However, the double-peak distribution feature observed here in our data implies that G4Ms may be an unfavorable element at the transcription initiation site.

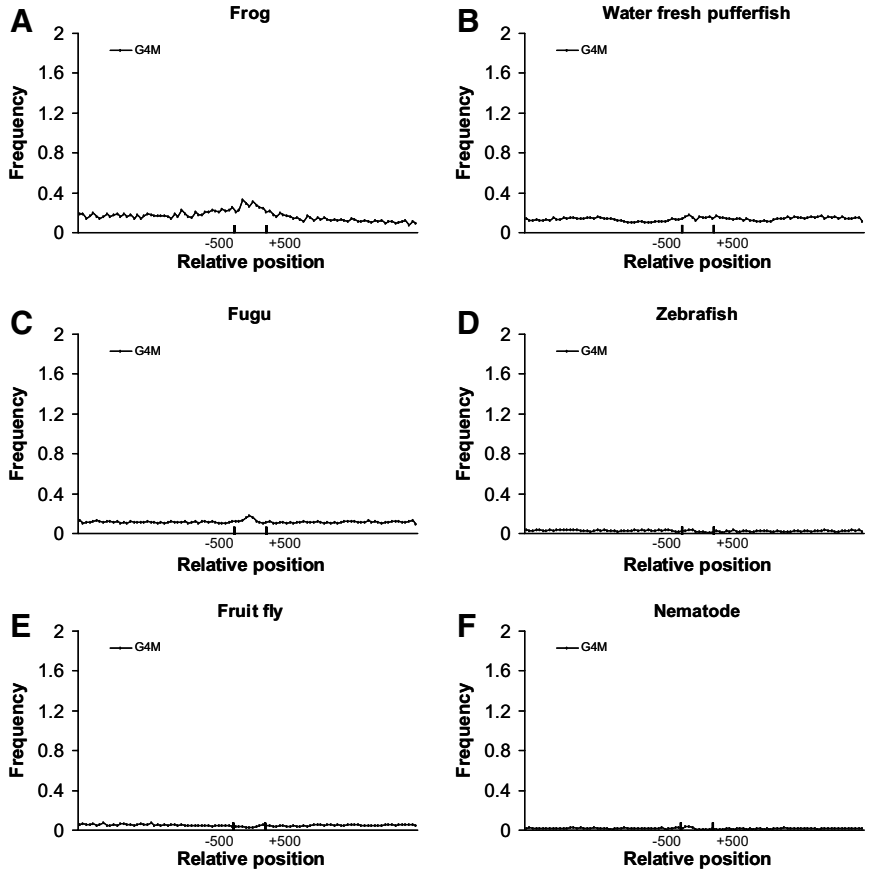


Fig. 2. Distribution of G4Ms in the TSS-flanking region of cold-blooded animals. Compared to warm-blooded animals, neither the TRR-enriched pattern of G4Ms nor the double-peak feature is detected in cold-blooded animals. Only the frequency of overall G4Ms is presented (black lines) because of the low frequency of G4Ms.

*3.3. Enrichment of G4Ms is not specific to warm-blooded genes*

The enrichment of the G4M gene in warm-blooded animals also raises the question as to whether or not the G4M gene is specific to warm-blooded animals. To address this, we classified the best-annotated human genes into six temporal groups (hereafter referred to as TG; see Section 2). The number of human genes classified into each temporal group (TG1 to TG6, from old to young) was 319, 3870, 2453, 3632, 7414 and 1088, respectively (Fig. 3A). The assumption was made that warm-blooded genes would be largely present in younger temporal groups. We analyzed the frequency of both overall, as well as stable G4Ms, along 10 kb TSS-flanking regions of human genes in each TG and, in addition, the corresponding orthologs in two selected lower species (fly and fugu).

As shown in Fig. 3B, both the TRR-enriched pattern of G4Ms and the double-peak feature were detected in the human genes of all the TGs. It was surprising that older human genes showed G4Ms richer than younger ones, when considering both the frequency of G4Ms and the ratio of the G4M gene (Fig. 3A and B), which was opposite to the expectation that the G4M gene was specific to warm-blooded animals. The ratio of the G4M gene in the two older temporal groups (TG1 and TG2) was 70.9% and 71.8%, whereas this decreased to 53.4% and 48.7% in the two younger TGs (TG5 and TG6).

It therefore appears that the selection for enrichment of G4Ms in the TRRs is not specific to warm-blooded genes, as G4Ms would have been adopted in the regulatory region of higher animals in a continuous manner. Indeed, orthologs of many older, G4M-negative or G4M-poor genes in cold-blooded animals (Figs. 2 and 3C) are becoming G4M-positive or G4M-rich in the human genome, and they are now even more G4M-rich than newly created ones (Fig. 3B). It should be noted, however, that although we provide evidence here that suggests selection for G4Ms is not warm-blood gene-specific, we do not preclude the possibility that different types of genes are under this kind of selection of different strengths. A recent example of functional preference of G4Ms is that proto-oncogenes and tumor repressors tend to have a very high and low frequency of G4Ms, respectively, in transcribed regions [26].

In summary, we have characterized the frequency distribution of putative G4Ms in the TSS-flanking regions across 13 animal genomes. We discovered two significant findings: (i) a systematic elevation of G4Ms exists in a wide range of genes in warm-blooded animals and (ii) overall G4Ms, as well as stable G4Ms are constitutively enriched in the TRRs of warm-blooded animals. In a living cell, most genomic DNA is maintained as a duplex and further packaged into chromatin, making the formation of G4 DNA more difficult than in vitro [2,6]. However, during the process of transcription, single-stranded DNA is generated locally, providing an opportunity for the formation of G4 DNA in the TRRs. Our findings support the hypothesis that G4Ms are involved in transcrip-

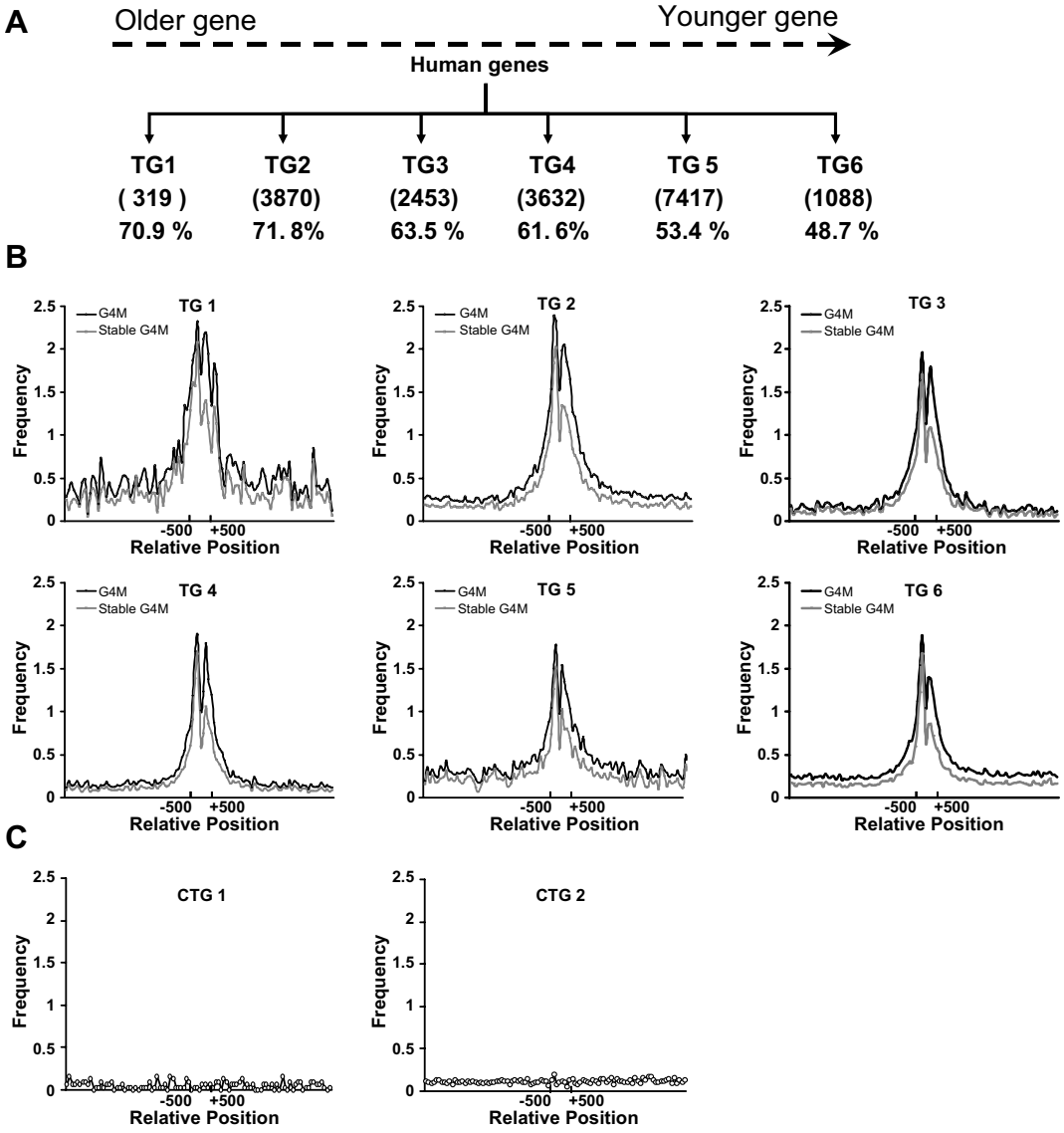


Fig. 3. Distribution of G4Ms in human genes in each temporal group and in corresponding fly and fugu orthologs. (A) Number of human genes classified into each TG; the corresponding ratio of G4M genes in each TG is listed below. (B) The frequency of overall and stable G4Ms (black and grey lines, respectively) in human genes in each TG. (C) The frequency of G4Ms in corresponding fly and fugu orthologs (CTG1 and CTG2).

tional regulation and also suggest that G4Ms are a novel type of regulatory motif that may contribute to the complexity of gene transcription in warm-blooded animals. At the same time, our results are also consistent with the notion that regulatory elements are also preferred by selection like gene coding regions [27,28].

*Acknowledgements:* This work was supported by the National Major Basic Research Program of China (973 program) and the National Natural Science Foundation of China. We thank the two anonymous reviewers for their constructive suggestions on our manuscript.

## References

[1] Simonsson, T. (2001) G-quadruplex DNA structures–variations on a theme. Biol. Chem. 382, 621–628.
[2] Han, H. and Hurley, L.H. (2000) G-quadruplex DNA: a potential target for anti-cancer drug design. Trends Pharmacol. Sci. 21, 136–142.
[3] 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.
[4] Keniry, M.A. (2000) Quadruplex structures in nucleic acids. Biopolymers 56, 123–146.
[5] Shafer, R.H. and Smirnov, I. (2000) Biological aspects of DNA/RNA quadruplexes. Biopolymers 56, 209–227.
[6] Maizels, N. (2006) Dynamic roles for G4 DNA in the biology of eukaryotic cells. Nat. Struct. Mol. Biol. 13, 1055–1059.
[7] Siddiqui-Jain, A., Grand, C.L., Bearss, D.J. and Hurley, L.H. (2002) Direct evidence for a G-quadruplex in a promoter region and its targeting with a small molecule to repress c-MYC transcription. Proc. Natl. Acad. Sci. USA 99, 11593–11598.
[8] 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.
[9] Rawal, P. et al. (2006) Genome-wide prediction of G4 DNA as regulatory motifs: role in Escherichia coli global regulation. Genome Res. 16, 644–655.
[10] Huppert, J.L. and Balasubramanian, S. (2007) G-quadruplexes in promoters throughout the human genome. Nucleic Acids Res. 35, 406–413.

[11] Grand, C.L., Powell, T.J., Nagle, R.B., Bearss, D.J., Tye, D., Gleason-Guzman, M. and Hurley, L.H. (2004) Mutations in the G-quadruplex silencer element and their relationship to c-MYC overexpression, NM23 repression, and therapeutic rescue. Proc. Natl. Acad. Sci. USA 101, 6140–6145.
[12] 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. USA 97, 12508–12512.
[13] 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.
[14] Dexheimer, T.S., Sun, D. and Hurley, L.H. (2006) Deconvoluting the structural and drug-recognition complexity of the G-quadruplex-forming region upstream of the bcl-2 P1 promoter. J. Am. Chem. Soc. 128, 5404–5415.
[15] 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.
[16] 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.
[17] 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.
[18] 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.
[19] Rankin, S. et al. (2005) Putative DNA quadruplex formation within the human c-kit oncogene. J. Am. Chem. Soc. 127, 10584–10589.
[20] Todd, A.K., Johnston, M. and Neidle, S. (2005) Highly prevalent putative quadruplex sequence motifs in human DNA. Nucleic Acids Res. 33, 2901–2907.
[21] Huppert, J.L. and Balasubramanian, S. (2005) Prevalence of quadruplexes in the human genome. Nucleic Acids Res. 33, 2908–2916.
[22] 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.
[23] Rivera, M.C., Jain, R., Moore, J.E. and Lake, J.A. (1998) Genomic evidence for two functionally distinct gene classes. Proc. Natl. Acad. Sci. USA 95, 6239–6244.
[24] Hazel, P., Huppert, J., Balasubramanian, S. and Neidle, S. (2004) Loop-length-dependent folding of G-quadruplexes. J. Am. Chem. Soc. 126, 16405–16415.
[25] Risitano, A. and Fox, K.R. (2004) Influence of loop size on the stability of intramolecular DNA quadruplexes. Nucleic Acids Res. 32, 2598–2606.
[26] Eddy, J. and Maizels, N. (2006) Gene function correlates with potential for G4 DNA formation in the human genome. Nucleic Acids Res. 34, 3887–3896.
[27] Gilad, Y., Oshlack, A., Smyth, G.K., Speed, T.P. and White, K.P. (2006) Expression profiling in primates reveals a rapid evolution of human transcription factors. Nature 440, 242–245.
[28] Nielsen, R. (2006) Comparative genomics: difference of expression. Nature 440, 161.