Published in final edited form as:
Mol Carcinog. 2009 April ; 48(4): 319–325. doi:10.1002/mc.20496.

Selection for the G4 DNA Motif at the 5' End of Human Genes

Johanna Eddy[1] and **Nancy Maizels**[1,2,*]

[1]Molecular and Cellular Biology Graduate Program, University of Washington School of Medicine, 1959 N.E. Pacific Street, Seattle, WA 98195-7650 USA

[2]Departments of Immunology and Biochemistry, University of Washington School of Medicine, 1959 N.E. Pacific Street, Seattle, WA 98195-7650 USA

Abstract

Formation of G4 DNA may occur in the course of replication and transcription, and contribute to genomic instability. We have quantitated abundance of G4 motifs and potential for G4 DNA formation of the nontemplate strand of 5’ exons and introns of transcripts of human genes. We find that, for all human genes, G4 motifs are enriched in 5’ regions of transcripts relative to downstream regions; and in 5’ regulatory regions relative to coding regions. Notably, although tumor suppressor genes are depleted and proto-oncogenes enriched in G4 motifs, abundance of G4 motifs in the 5’ regions of transcripts of genes in these categories do not differ. These results support the hypothesis that G4 motifs are under selection in the human genome. They further show that for tumor suppressor genes and proto-oncogenes, independent selection determines G4 DNA potential of 5’ regulatory regions of transcripts and downstream coding regions.

Keywords

intron; quadruplex; oncogene; tumor suppressor gene

INTRODUCTION

DNA has considerable potential to form structures other than canonical Watson-Crick duplexes, with potentially profound consequences for genomic stability. One such structure, G4 DNA, forms readily in G-rich sequences and is unusually stable once formed [1–3]. The essential unit of G4 DNA is the G-quartet, a planar array of four guanines in which each guanine pairs with two neighbors by hydrogen bonding. G4 DNA formation requires that a sequence contain at least four runs of guanines, each at least three nt in length; thus the G4 signature motif is $G_3N_xG_3N_xG_3N_xG_3$. Transient denaturation accompanying replication or transcription enables G4 DNA formation (Figure 1). Because G4 DNA forms transiently, it is difficult to detect directly in a living cell.

G-quartets and G4 structures can also form in RNA, in sequences containing the G4 motif {Patel, 2007 #5076}. Sequences that form alternative structures (reviewers say H-DNA, triplex, etc) could also be G-rich and contain G4 motifs, and these DNA structures may also be important ??? However, the discovery of factors that recognize G4 DNA structures suggests that G4 DNA is an alternate structure that forms *in vivo*. Nancy – help! I was thinking of adding reviewer concern about other structures here, but maybe the Discussion is better? And I definitely need help knowing what to say as I am not very familiar with other structures.

[*]To whom correspondence should be addressed. Tel 206 221 6876; Fax 206 221 6781; maizels@u.washington.edu.

Deficiencies in factors that recognize or resolve G4 DNA structures cause genomic instability leading to cancer. The list of such factors currently includes the RecQ family helicases, BLM [4] and WRN [5]; the SF2 (RAD3) family helicase, FANCJ, deficient in Fanconi anemia [6]; and the mismatch repair factor MutSα [7]. Human BLM and FANCJ helicases are both associated with the replication apparatus [8] and they unwind G4 DNA *in vitro* with opposite polarities [4,6] . Genetic analyses in *C. elegans* has shown that the nematode FANCJ homolog, DOG-1, is essential for stabilization of genomic regions that bear the G4 signature motif; and that this instability is enhanced by the absence of the nematode BLM homolog, HIM-6 [9]. Thus, replicative stability of regions bearing the G4 motif may depend primarily on the 5’-3’ helicase, FANCJ, while the 3’–5’ helicase BLM provides a secondary pathway [10].

Genomic instability appears to be inherent to regions bearing the G4 signature, and this may reflect specialized requirements for replication. Some of the most variable of the “variable number tandem repeats” (VNTRs) in the human genome carry the G4 signature motif, such as the MS1 minisatellite (D1S7), AGGGTGGAG; D4S43, GGGGAGGGGGAAGA; and the insulin-linked hypervariable repeat, ACAGGGGTGTGGGG [11]. This motif is also found at sites of instability leading to genetic disease, such as the 22q13.3 deletion resulting in mental retardation [12]; and the CGCGGGGCGGGG repeat associated with progressive myoclonus epilepsy, type1 [13].

Transcription of regions bearing the G4 motif in the nontemplate strand results in formation of an unusual structure, the G-loop (Figure 1B). G-loops contain a stable cotranscriptional RNA/DNA hybrid on the template strand, and G4 DNA interspersed with single-stranded regions on the G-rich nontemplate strand [14–16]. Analysis of these structures at single molecule resolution showed that they can extend over hundreds of bp. G-loops form in transcribed G-rich regions including proto-oncogenes, immunoglobulin (Ig) switch regions and telomeric repeats. Stable RNA/DNA hybrids, like those within G-loops, are targets of genomic instability in eukaryotic cells, where factors associated with RNA processing combat their persistence [17,18].

The immune system appears to take advantage of the inherent instability of regions bearing the G4 motif in the process of class switch recombination, which joins a new downstream constant region to the expressed heavy chain variable region, thus changing antigen clearance properties of an immunoglobulin molecule without affecting antigen recognition. Class switch recombination targets recombination junctions to G-rich “switch” or S regions at the IgH heavy chain locus, which are transcribed from dedicated promoters to initiate and target switch recombination [19]. The B cell-specific enzyme, Activation-Induced Deaminase (AID), then deaminates cytosine to uracil, which undergoes mutagenic repair to create DNA breaks. AID preferentially attacks single-stranded DNA, and could thus target single-stranded regions in the nontemplate strand of a G-loop within a transcribed S region to promote genomic instability. This same mechanism may be the source of recurrent translocations characteristic of B cell tumors. Diffuse large B-cell lymphoma (DLBCL), the most common lymphoid malignancy in adults, derives from germinal center B cells which express AID. These tumors are characterized by recurrent translocations of c-*MYC* and other proto-oncogenes, which can contribute to initial tumor development and correlate with poor prognosis and aggressive tumor growth [20]. The recurrent translocations in AID-expressing tumors map to regions which are G-rich and form G-loops upon transcription [15,16].

The connection between the G4 motif and genomic instability has motivated recent analyses designed to learn where in the genome G4 DNA can form, and how it may contribute to gene regulation or gene function. These analyses have used software, developed by our own and other laboratories, to identify the G4 motif in complex genomes. Algorithms typically search for at least 4 runs of at least 3 guanines within a window ranging from approximately 30 to

100 nt in length. The upper limit has not been experimentally examined, and is somewhat arbitrary in current algorithms. The G4 signature motif, $G_3N_xG_3N_xG_3N_xG_3$, is readily recognized at the telomeres, rDNA, and Ig heavy chain switch regions; and at more than 300,000 distinct sites in the human genome [21].

Genomic analyses support the view that the G4 motif has been selected and counterselected in the course of evolution. Among single copy genes, the G4 motif correlates with gene function as defined by GO (gene ontology) terms: it is depleted in tumor suppressor genes, which maintain genomic stability and may exhibit haploinsufficiency [22]; but enriched in proto-oncogenes, which promote cell proliferation [23]. Promoters in humans and other vertebrates are enriched in G4 signature motifs [24–26]. We demonstrated that conserved motifs that function in duplex DNA (CpG dinucleotides, motifs for transcription factors such as Sp1) account for most matches to the G4 motif upstream of the transcription start site (TSS) in human promoters. However, nearly half of human genes contain elements that match the G4 signature downstream of the TSS, in the nontemplate strand at the very 5' end of intron 1; and these elements are robust to masking motifs, such CpG dinucleotides, with defined functions in duplex DNA. These G-rich intron 1 (GrIn1) elements are conserved through fish. There is a clear strand bias, which would enable G4 structures to form in either DNA or RNA.

The position of GrIn1 elements at the very 5’ end of the first intron would enable them to regulate events at or near the promoter. This raised the question of whether G4 motifs might be enriched other regions of transcripts. We have therefore quantitated G4 DNA potential (G4P) of the nontemplate strand of 5’ exons and introns of transcripts of human genes, separately examining exons that do or do not have coding capacity. We find that, for all human genes, G4 motifs are enriched in 5’ regions of transcripts relative to downstream regions. Notably, although tumor suppressor genes are depleted and proto-oncogenes enriched in G4 motifs, abundance of G4 motifs in the 5’ regions of transcripts of genes in these categories do not differ. These results support the hypothesis that G4 motifs are under selection, and strongly suggest that independent selection determines G4P of 5’ regulatory regions of transcripts and downstream coding regions.

MATERIALS AND METHODS

Human gene sequence data (NCBI 36 assembly) and exon definitions, including sequence type (coding, 5’ UTR, or 5’ UTR and coding), were obtained from Ensembl 46, using BioMart [28,29]. We analyzed 194,951 exon sequences with unique Ensembl Exon IDs from all transcripts of all human RefSeq genes. The 55 tumor suppressor genes and 95 proto-oncogenes were previously defined [23]. Intron sequences were derived from transcript sequences as previously described [27].

G4 DNA potential was calculated with the “G4P Calculator” program [23], available on our web site (http://depts.washington.edu/maizels9), setting window size to 100 nt, window shift to 1 nt, minimum size of G-run to 3, and minimum number of G-runs per window to 4. Our previous analysis used a window shift of 20 nt when analyzing the longer complete gene sequences [23]. G4P is calculated as the number of windows that meet the defined criteria divided by the number of windows searched. For sequences that are shorter than 100 nt, such as many exons, G4P calculated by this method may be artificially high. Therefore we normalized G4P by sequence length, designated as G4p. For example, a 75 bp sequence with one G4 motif has G4P = 100% and G4p = 1.3%.

G4p comparisons were done by two-tailed, unpaired, Wilcoxon rank-sum test using the statistics program R v2.7.1, and significance determined by $P < 0.01$.

RESULTS

G4 DNA Motifs are Depleted in Coding Exons

We first evaluated G4 DNA potential (G4P) of the nontemplate strand of 194,951 unique exon sequences from all human RefSeq genes (NCBI 36), using G4P Calculator [23]. This tool quantitates G4P by searching for matches to the G4 signature motif within overlapping windows of sequence. We normalized G4P per sequence length, designated as G4p, to permit accurate comparison independent of length, as otherwise short sequences could bias quantitation. This analysis showed that most exons (85%) have no potential to form G4 DNA (G4p = 0). Median G4p is zero; and average G4p = 2.5% (Table 1).

Most exons contain coding sequence only. For 93% of these coding exons, G4p = 0; and for all coding exons, average G4p = 1.1%. Among all other exons (including 5’ UTR, 3’ UTR, 5’ and 3’ UTR, or unclassified), 61% have G4p = 0; while 49% have G4p > 0; and average G4p = 6.7%. The difference between G4p of coding and other exons is significant ($P << 10^{-16}$, Wilcoxon test), so G4 motifs are depleted in coding exons; and exons containing regulatory sequence may be under selection distinct from exons containing coding sequence.

G4p of the Nontemplate Strand is Low in Tumor Suppressor Genes and High in Proto-oncogenes

We previously showed that tumor suppressor genes are depleted (median G4P = 2.4%) and proto-oncogenes enriched (median G4P = 11.0%) in G4 motifs on the nontemplate strand relative to all RefSeq genes (median G4P = 5.0%) [23]. Similar differences are evident upon calculation of G4p. Quantitation of G4p of the 55 tumor suppressor genes and 95 proto-oncogenes (NCBI 36) showed that the average nontemplate strand G4p of tumor suppressor genes (4.2%) was lower than that of the RefSeq genes (9.1%; NCBI 34), while that of the proto-oncogenes (12.7%) was higher than that of the RefSeq genes (Figure 2). This difference between tumor suppressor genes and proto-oncogenes was significant ($P = 5.4 \times 10^{-9}$).

G4p of Coding Exons Correlates with Gene Classification

To establish whether differences in G4p of tumor suppressor and proto-oncogenes could be ascribed to differences in coding exons, G4p of the nontemplate strand of coding exons was examined separately. Because most exons have G4p = 0 (Table 1), we compared average G4p rather than median G4p. This analysis showed that the average nontemplate strand G4p of tumor suppressor gene coding exons (0.5%) was lower than that of RefSeq gene coding exons (1.1%); while this average for proto-oncogene coding exons (1.7%) was higher than for RefSeq genes (Figure 2). The difference in coding exon G4p between tumor suppressor genes and proto-oncogenes was significant ($P = 2.3 \times 10^{-9}$). Thus, differences in G4p of coding exons contribute to the differences between G4p of these genes.

G4p of 5’ Noncoding Exons Does Not Correlate with Gene Classification

Exons at the very 5’ end of transcripts can be classified by whether they contain coding sequence, regulatory sequence (ATG translation start sites, 5’ UTR) or both (Figure 3A). To ask if G4p at the 5’ end of transcripts correlated with G4p of genes, we examined G4p of the nontemplate strand by exon type for tumor suppressor genes and proto-oncogenes. The average G4p of exons containing only 5’ UTR sequences was identical for tumor suppressor and proto-oncogenes (5.7%, $P = 0.31$; Figure 3B), and similar to that of all RefSeq genes (5.2%). Comparison of average nontemplate strand G4p of exons containing both 5’ UTR and coding sequence (including ATG) showed that this value was somewhat higher for proto-oncogenes (11.0%) than for all RefSeq genes (8.8%), and lower for tumor suppressor genes (8.2%; Figure 3B); however, the difference between tumor suppressor and proto-oncogenes was not significant (*P* = 0.10). The absence of difference in G4p of 5’ noncoding exons is in contrast to the very significant difference between G4p of coding exons. This result strongly suggests that G4p of the 5’ noncoding exons is under selection independent of G4p of genes.

G-richness at the 5’ End of the First Intron Does Not Correlate with Gene Classification

We previously found that the 5’ end of the nontemplate strand of the first intron of many human genes is more G-rich than the genomic average, suggesting that a G-rich regulatory element might be located within this intron [27]. To ask if this feature also characterized tumor suppressor and/or proto-oncogenes, we quantitated G4p of the nontemplate strand of first introns of genes in these classes. We found that the average G4p of first introns of tumor suppressor genes was 13.2%, comparable to the average G4p of the first intron of all RefSeq genes (12.8%; Figure 4). Average G4p of first introns of the proto-oncogenes was even higher (17.0%). The difference between G4p of first introns of tumor suppressor and proto-oncogenes was significant (*P* = 0.0008). To ask if G4p of the first intron differs from G4p of downstream introns, we calculated G4p of second and third introns. The average G4p of second introns of tumor suppressor genes was 3.3%, and that of the proto-oncogenes was 12.4% ($P = 1.5 \times 10^{-7}$); while that of RefSeq genes was 8.2% (Figure 4A). The average G4p of third introns of tumor suppressor genes was 3.0%, and that of proto-oncogenes was 11.9% ($P = 2.3 \times 10^{-5}$); while that of RefSeq genes was 7.6%. Moreover, average G4p values of intron 2 and intron 3 were similar to average G4p of the complete gene sequences (tumor suppressor genes, 4.2%; proto-oncogenes, 12.7%; compare Figure 4A and Figure 2). This is not surprising, given that, for most genes, intron sequences are a major component of the entire gene sequence.

Since the difference between tumor suppressor genes and proto-oncogenes in average G4p of the first intron is not as great as in the second and third introns, and since we previously found that almost half of all genes contain a G4 motif at the 5’ end of the first intron {Eddy, 2008 #5203}, we took a closer look at the 5’ region of the first introns. We plotted the average number of G-runs on the nontemplate strand, in 100 nt intervals, for 1000 nt of the first introns (Figure 4B), and compared tumor suppressor genes to proto-oncogenes at each interval. At the 5’ end, the average number of G-runs for tumor suppressor genes is 3.3 and 43% of the introns have a G4 motif (4 or more G-runs), in comparison to 48% of the RefSeq genes. The average number of G-runs for proto-oncogenes is 3.9, and 50% of the introns have a G4 motif. The difference between the tumor suppressor gene and proto-oncogene groups is not significant in the two 5’-most intervals (*P* = 0.2 and 0.4), but is significant (*P* < 0.01) in the downstream intervals.

The evidence that potential for G4 DNA formation at the 5’ end of first introns of tumor suppressor genes and proto-oncogenes is not significantly different, despite the contrasting G4p of genes in these categories, supports the notion that G4p of the 5’ end of the first intron is under selection independent of G4p of other introns, or of the genes themselves.

DISCUSSION

We have shown that the G4 motif is nonrandomly distributed in human genes. By genomic analysis of the nontemplate strand, we have shown that the G4 motif is enriched in 5’ exons containing regulatory sequences, either 5’ UTR alone or 5’ UTR and ATG motifs; and depleted in coding exons. The G4 motif is also enriched in first introns, compared to downstream introns {Eddy, 2008 #5203}. Exons containing 5’ UTR sequence are likely to contribute to regulation of gene expression. The enrichment of G4p in these exons is consistent with the notion that G4 motifs contribute to regulation of gene expression. The mechanism of such regulation has yet to be defined.

The possibility that G4 motifs are under selection was further supported by analysis of tumor suppressor genes and proto-oncogenes, which are depleted and enriched for G4p, respectively [23]. In these two classes of genes, the G4p of coding exons was low for tumor suppressor genes and high for proto-oncogenes, exhibiting the same relationship as in the gene sequences. The significant difference between G4P of coding exons of tumor suppressor genes and proto-oncogenes is noteworthy. Sequences of coding exons are selected for a functional protein product, demands of translation are thought to dominate selection. This evidence for additional selection based on G4p now shows that G4p may be important for gene or genome function.

High G4p in the 5' regulatory regions was not correlated with G4p of the coding regions, as it was found associated with genes for which the complete transcribed sequence is characterized by either high or low G4p. This further supports the notion that G4p of 5' regulatory regions is under selection distinct from selection that determines G4p of genes.

Despite the differences between G4p in tumor suppressor genes and proto-oncogenes, there was not a significant difference in G-richness at the 5' end of first introns of genes in these categories. Thus, for these two functional classes of genes, the 5' end of the first intron is under selection independent of G4p of other introns, or of the genes themselves. Independent selection of G-rich sequence at the 5' end of the first intron was anticipated by our discovery that almost half of all human genes contain a G4 motif in this region [27].

These results provide further support for the hypothesis that the G4 motif is selected in the human genome. They also show that both tumor suppressor and proto-oncogenes have been under independent selection for G4p of the genes themselves and G4p of 5'-regulatory regions of the transcript, including 5' UTR, first exon and first intron sequences. Selection for these sequences could be at the level of DNA or RNA. Whether independent selection similarly governs G4p of other genes and 5' regulatory regions of their transcripts is an open question for future research.

Acknowledgments

Funding was provided by National Institutes of Health R01 GM065988 and P01 CA077852 to NM; National Cancer Institute CA009537 (Basic and Cancer Immunology Training Grant) and Cancer Research Institute (Tumor Immunology Pre-doctoral Training Grant) to JE.

REFERENCES

1. Sen D, Gilbert W. Formation of parallel four-stranded complexes by guanine rich motifs in DNA and its implications for meiosis. Nature 1988;334:364–366. [PubMed: 3393228]
2. Maizels N. Dynamic roles for G4 DNA in the biology of eukaryotic cells. Nat Struct Mol Biol 2006;13:1055–1059. [PubMed: 17146462]
3. Phan AT, Kuryavyi V, Patel DJ. DNA architecture: from G to Z. Curr Opin Struct Biol 2006;16:288–298. [PubMed: 16714104]
4. Sun H, Karow JK, Hickson ID, Maizels N. The Bloom's syndrome helicase unwinds G4 DNA. J Biol Chem 1998;273:27587–27592. [PubMed: 9765292]
5. Bachrati CZ, Hickson ID. Analysis of the DNA unwinding activity of RecQ family helicases. Methods Enzymol 2006;409:86–100. [PubMed: 16793396]
6. Wu Y, Shin-Ya K, Brosh RM Jr. FANCJ helicase defective in Fanconi anemia and breast cancer unwinds G-quadruplex DNA to defend genomic stability. Mol Cell Biol. 2008
7. Larson ED, Duquette ML, Cummings WJ, Streiff RJ, Maizels N. MutSalpha binds to and promotes synapsis of transcriptionally activated immunoglobulin switch regions. Curr Biol 2005;15:470–474. [PubMed: 15753043]
8. Wang W. Emergence of a DNA-damage response network consisting of Fanconi anaemia and BRCA proteins. Nat Rev Genet 2007;8:735–748. [PubMed: 17768402]

9. Youds JL, O'Neil NJ, Rose AM. Homologous recombination is required for genome stability in the absence of DOG-1 in Caenorhabditis elegans. Genetics 2006;173:697–708. [PubMed: 16547095]

10. Maizels N. Genomic Stability: FANCJ-Dependent G4 DNA Repair. Curr Biol 2008;18:R613–R614. [PubMed: 18644339]

11. Weitzmann MN, Woodford KJ, Usdin K. DNA secondary structures and the evolution of hypervariable tandem arrays. J Biol Chem 1997;272:9517–9523. [PubMed: 9083093]

12. Bonaglia MC, Giorda R, Mani E, et al. Identification of a recurrent breakpoint within the SHANK3 gene in the 22q13.3 deletion syndrome. J Med Genet 2006;43:822–828. [PubMed: 16284256]

13. Saha T, Usdin K. Tetraplex formation by the progressive myoclonus epilepsy type-1 repeat: implications for instability in the repeat expansion diseases. FEBS Lett 2001;491:184–187. [PubMed: 11240124]

14. Duquette ML, Handa P, Vincent JA, Taylor AF, Maizels N. Intracellular transcription of G-rich DNAs induces formation of G-loops, novel structures containing G4 DNA. Genes Dev 2004;18:1618–1629. [PubMed: 15231739]

15. Duquette ML, Pham P, Goodman MF, Maizels N. AID binds to transcription-induced structures in c-MYC that map to regions associated with translocation and hypermutation. Oncogene 2005;24:5791–5798. [PubMed: 15940261]

16. Duquette ML, Huber MD, Maizels N. G-rich proto-oncogenes are targeted for genomic instability in B-cell lymphomas. Cancer Res 2007;67:2586–2594. [PubMed: 17363577]

17. Aguilera A. mRNA processing and genomic instability. Nat Struct Mol Biol 2005;12:737–738. [PubMed: 16142225]

18. Li X, Manley JL. Cotranscriptional processes and their influence on genome stability. Genes Dev 2006;20:1838–1847. [PubMed: 16847344]

19. Maizels N. Immunoglobulin gene diversification. Annu Rev Genet 2005;39:23–46. [PubMed: 16285851]

20. Jankovic M, Nussenzweig A, Nussenzweig MC. Antigen receptor diversification and chromosome translocations. Nat Immunol 2007;8:801–808. [PubMed: 17641661]

21. Huppert JL, Balasubramanian S. Prevalence of quadruplexes in the human genome. Nucleic Acids Res 2005;33:2908–2916. [PubMed: 15914667]

22. Payne SR, Kemp CJ. Tumor suppressor genetics. Carcinogenesis 2005;26:2031–2045. [PubMed: 16150895]

23. Eddy J, Maizels N. Gene function correlates with potential for G4 DNA formation in the human genome. Nucleic Acids Res 2006;34:3887–3896. [PubMed: 16914419]

24. Huppert JL, Balasubramanian S. G-quadruplexes in promoters throughout the human genome. Nucleic Acids Res 2007;35:406–413. [PubMed: 17169996]

25. Du Z, Kong P, Gao Y, Li N. Enrichment of G4 DNA motif in transcriptional regulatory region of chicken genome. Biochem Biophys Res Commun 2007;354:1067–1070. [PubMed: 17275786]

26. Zhao Y, Du Z, Li N. Extensive selection for the enrichment of G4 DNA motifs in transcriptional regulatory regions of warm blooded animals. FEBS Lett 2007;581:1951–1956. [PubMed: 17462634]

27. Eddy J, Maizels N. Conserved elements with potential to form polymorphic G-quadruplex structures in the first intron of human genes. Nucleic Acids Res 2008;36:1321–1333. [PubMed: 18187510]

28. Hubbard TJ, Aken BL, Beal K, et al. Ensembl 2007. Nucleic Acids Res 2007;35:D610–D617. [PubMed: 17148474]

29. Durinck S, Moreau Y, Kasprzyk A, et al. BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. Bioinformatics 2005;21:3439–3440. [PubMed: 16082012]

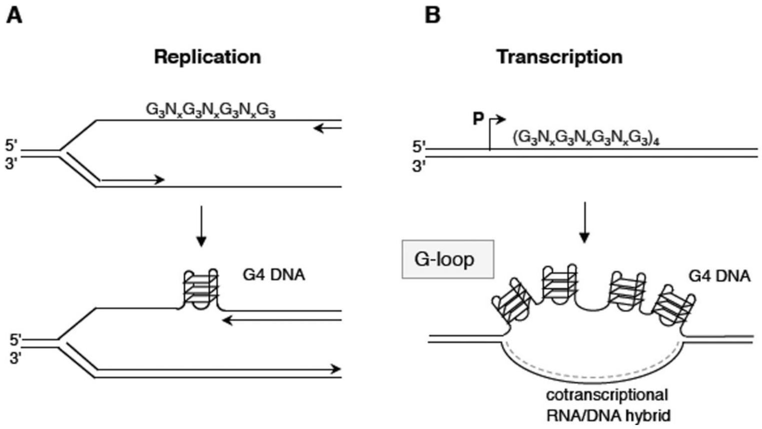

Figure 1.

Formation of G4 DNA during replication or transcription of regions bearing the G4 motif. For purposes of illustration, regions are shown with (A) one or (B) four matches to the G4 signature.

(A) G4 DNA formation during replication (shown to occur on the lagging strand).

(B) Transcription results in formation of a G-loop, which contains a stable cotranscriptional RNA/DNA hybrid (dotted gray line) and G4 DNA.

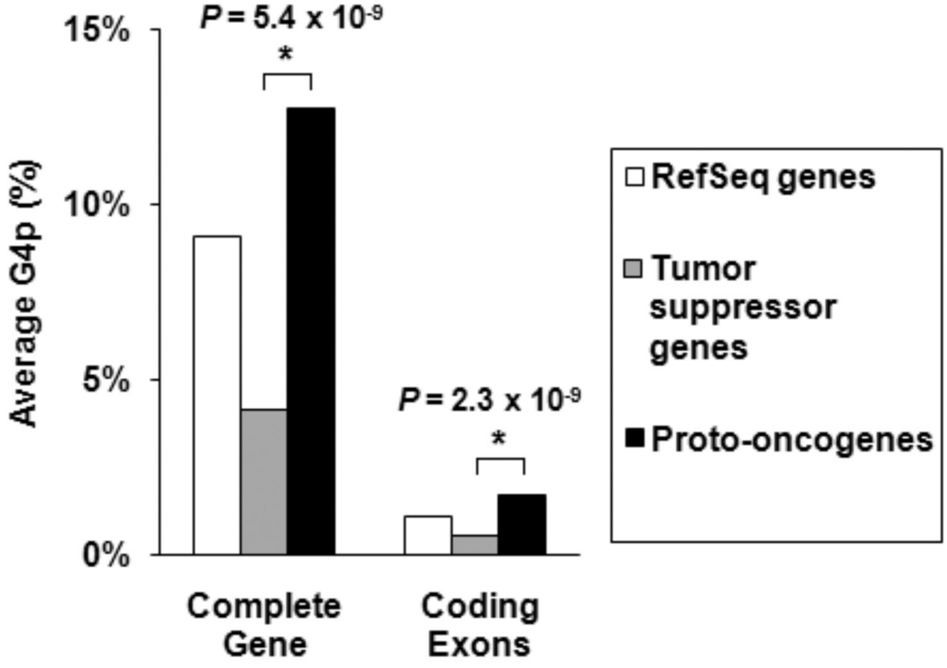

Figure 2.

G4p of the nontemplate strand and of coding regions correlates with gene classification. Average G4p of complete genes and coding exons only. G4p was calculated for all RefSeq genes (>16,000 genes, NCBI 34, white), 55 tumor suppressor genes (NCBI 36, gray), and 95 proto-oncogenes (NCBI 36, black). Asterisks mark significant differences between tumor suppressor gene and proto-oncogene sequences, measured by Wilcoxon rank-sum test.

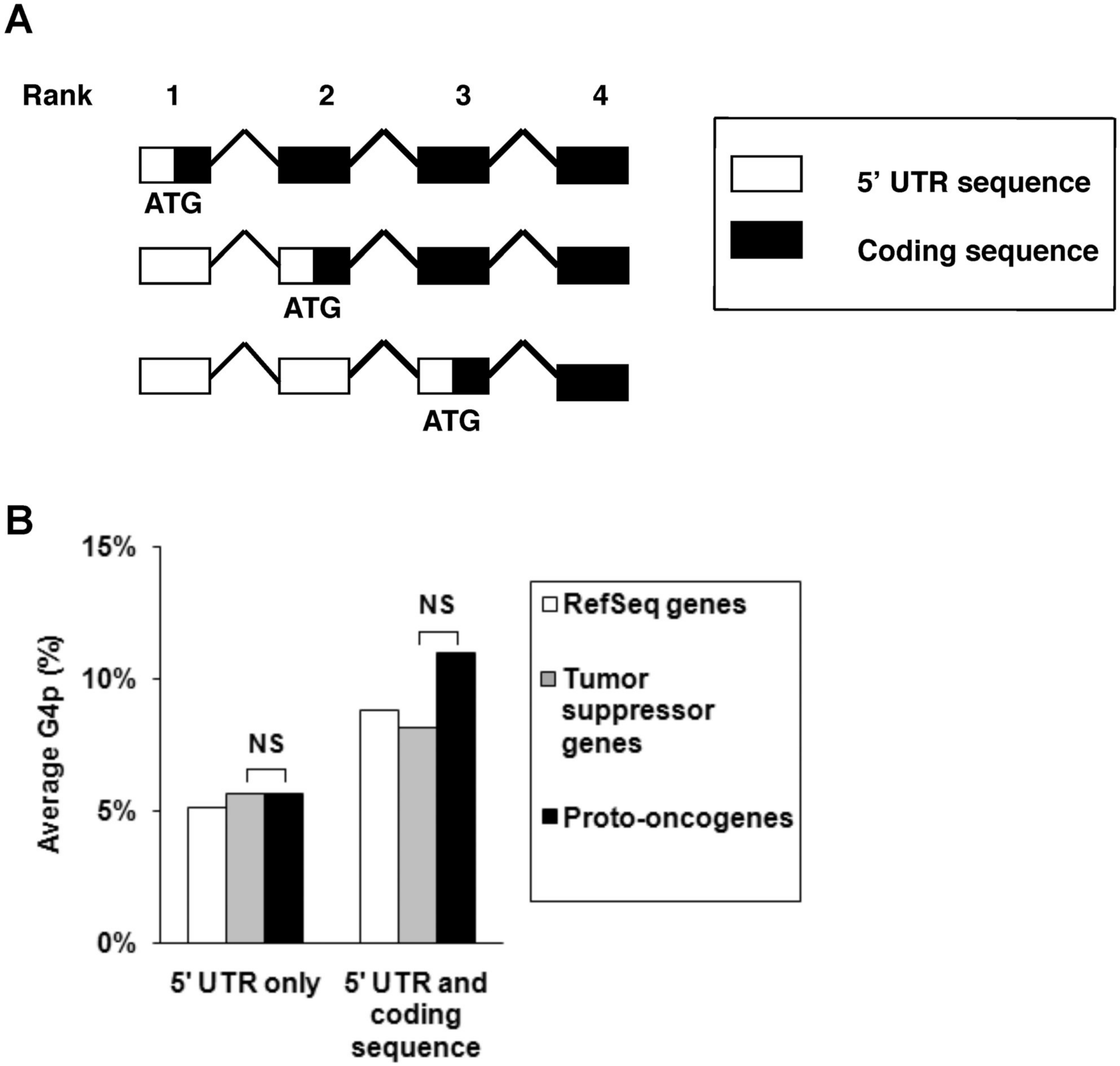

Figure 3.

G4p of 5’ noncoding exons does not correlate with gene classification.

(A) Diagram of gene organization and classification of 5’ exon types, showing 5’ UTR (white), coding sequences (black); and exons containing the ATG translation start site (and thus both 5' UTR and coding sequences, white and black). Note that ATG start sites may be in first, second or third exons, as shown, but these were not separately analyzed.

(B) Average G4p of exons containing only 5’ UTR sequence (white boxes in panel A), or both 5’ UTR sequence and coding sequence (white and black boxes in panel A). G4p was calculated for all RefSeq genes (white), tumor suppressor genes (gray), and proto-oncogenes (black). NS indicates that differences between tumor suppressor gene and proto-oncogene sequences were not significant, as measured by Wilcoxon rank-sum test.

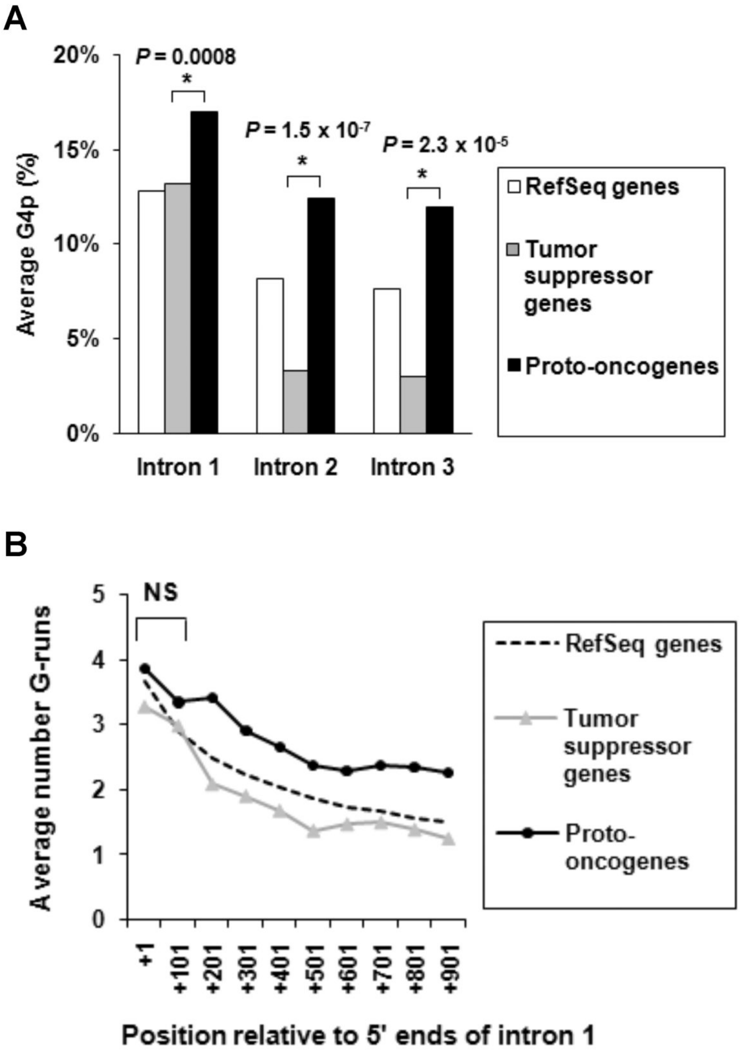

Figure 4.

G-richness at the 5’ end of the first intron does not correlate with gene classification.

(A) Average G4p of first, second or third introns. G4p was calculated for all RefSeq genes (white), tumor suppressor genes (gray), and proto-oncogenes (black). Asterisks mark significant differences between tumor suppressor gene and proto-oncogene sequences, measured by Wilcoxon rank-sum test.

(B) Average number of G-runs per 100 nt in first introns. G-runs per 100 nt for 1000 nt of first introns were counted for RefSeq genes (dotted line), tumor suppressor genes (gray line), and proto-oncogenes (black line). NS indicates intervals that did not show significant differences between tumor suppressor gene and proto-oncogene sequences, as measured by Wilcoxon rank-sum test. Unmarked intervals were significantly different (*P* < 0.01).

Table 1

G4p of exons.

	All Exons	Coding Exons	Noncoding Exons
Number of exons	194,951	147,062	47,889
% of exons with G4p > 0	15%	7%	39%
% of exons with G4p = 0	85%	93%	61%
Average G4p (%)	2.5%	1.1%	6.7%