Nucleic Acids Research Advance Access published August 11, 2006

Nucleic Acids Research, 2006, Vol. 00, No. 00 ***1–10***
doi:10.1093/nar/gkl529

Gene function correlates with potential for G4 DNA formation in the human genome

Johanna Eddy[1] and Nancy Maizels[1,2,3,*]

[1]Molecular and Cellular Biology Graduate Program, [2]Department of Immunology and [3]Department of Biochemistry, University of Washington School of Medicine, 1959 NE Pacific Street, Seattle, WA 98195-7650, USA

Received June 13, 2006; Revised July 9, 2006; Accepted July 10, 2006

ABSTRACT

G-rich genomic regions can form G4 DNA upon transcription or replication. We have quantified the potential for G4 DNA formation (G4P) of the 16 654 genes in the human RefSeq database, and then correlated gene function with G4P. We have found that very low and very high G4P correlates with specific functional classes of genes. Notably, tumor suppressor genes have very low G4P and proto-oncogenes have very high G4P. G4P of these genes is evenly distributed between exons and introns, and it does not reflect enrichment for CpG islands or local chromosomal environment. These results show that genomic structure undergoes selection based on gene function. Selection based on G4P could promote genomic stability (or instability) of specific classes of genes; or reflect mechanisms for global regulation of gene expression.

INTRODUCTION

Eukaryotic genomes contain characteristically G-rich regions, including single-copy genes; the rDNA; and repetitive sequences, such as the telomeres and the immunoglobulin heavy chain switch (S) regions of higher vertebrates. G-rich nucleic acids have the potential to form G-quadruplex or 'G4 DNA', a structure in which intra- or inter-strand interactions are stabilized by G-quartets, planar arrays of four guanines, paired by Hoogsteen bonding (1,2). G-quartets can stabilize a remarkable diversity of structures, in which the lengths and positions of the G-runs and the 'loops' separating them both contribute to overall topology (3,4). In the human genome, the number of distinct sites with potential to form G4 DNA is estimated at more than 300 000, and specific loop sequences are prominent at some of these sites (5,6).

Key cellular processes are identified with repetitive G-rich chromosomal regions, where regulated formation of G4 DNA may contribute to biological function. At the G-rich telomere tails, the presence of G4 DNA inhibits extension by telomerase, and proteins that bind specifically to telomeric sequences regulate the formation and resolution of G4 DNA (7–15). The G-rich immunoglobulin switch regions are sites of recombination that is critical to B-cell development and the immune response, and regulated transcription of the switch regions induces the formation of DNA structures targeted by factors essential to class switch recombination (16,17).

G-rich regions can also be sites of unprogrammed genomic instability. Many B-cell lymphomas carry a translocation of the *MYC* proto-oncogene to the immunoglobulin heavy chain switch region (18), and the common translocation breakpoints map to G-rich regions of *MYC* that form structures similar to those formed by transcribed G-rich switch regions (19,20). Some of the most unstable human minisatellites are G-rich sequences predicted to form G4 DNA (21); and G4 DNA formation *in vitro* has been directly confirmed for two G-rich VNTRs, D4S43, and the insulin-linked hypervariable repeat (22). Reporter constructs carrying interstitial telomeric repeats display high levels of instability (23), which may be analogous to the instability of G-rich VNTRs.

Specialized mechanisms may regulate the expression of G-rich genes at the levels of transcription, RNA processing and translation. Cotranscriptional RNA:DNA hybrid formation occurs readily within G-rich regions (19,24,25). Factors associated with RNA processing pathways, including THO/TREX and ASF/SF2, normally prevent cotranscriptional RNA:DNA hybrid formation, and promote gene expression; and genomic instability ensues in their absence (26,27). Factors involved in translational regulation may target RNA transcripts that contain G-quartets (28,29). Regions with the potential to form G4 DNA have been identified in the promoters of several proto-oncogenes, including c-*MYC*, *VEGF*, c-*KIT* and *BCL2* (30–33). This has led to suggestions that formation or resolution of specific quadruplex structures may contribute to the regulation of gene expression, and prompted the design of therapeutics targeted to these structures, but the biological specificity of such compounds is yet to be established rigorously (34–39).

Conserved and ubiquitous repair factors recognize G4 DNA, including the human RecQ family helicases BLM and WRN (40,41); the *Saccharomyces cerevisiae* RecQ family helicase Sgs1 (42); and the mismatch repair factor MutSα, a heterodimer of MSH2/MSH6 (16). RecQ family

*To whom correspondence should be addressed. Tel: +1 206 221 6876; Fax: +1 206 221 6781; Email: maizels@u.washington.edu

helicases maintain G-rich regions during replication. Sgs1 is required for nucleolar stability and replication of the G-rich rDNA (43,44); and in the absence of WRN helicase, telomeric sequence is lost due to impaired replication of the G-rich strand (45). The mismatch repair factor, MutSα, may cooperate with BLM helicase to promote the resolution of G4 DNA during replication (46). In immunoglobulin switch recombination, MutSα recognizes G4 DNA formed during transcription of the G-rich switch regions to promote their synapsis and recombination (16).

Genomic regions with potential to form G4 DNA have been enumerated (4,5), but they have not been correlated with specific gene functions. The link between potential for G4 DNA formation and genomic instability suggests that the identification of human genes with relatively high or low potential to form G4 DNA might provide insights into the evolution of genomic structure, or identify mechanisms that could account for genomic instability in human malignancies. The possibility that G-richness can contribute to shared regulation suggests that genes with similar or related functions may share features of genomic structure. We therefore set out to determine the prevalence of G-rich sequences capable of forming G4 DNA among human genes, and to determine if particular functional classes of genes might be characterized by the presence or absence of G-rich regions.

METHODS

G4P Calculator software

We developed a software program, 'G4P Calculator', which computes G4 DNA potential based on the density of runs of guanines in a sequence. The program evaluates runs of guanines in a sliding window and calculates the percentage of windows searched that meet the specified criteria. The criteria used are as follows: G-run length, $\geqslant 3$; number of G-runs per window, $\geqslant 4$; window length, 100 nt; and sliding interval length, 20 nt. The requirement for four or more runs of three or more guanines is based on studies of oligonucleotide folding [reviewed in (2)]. The 100 nt window size facilitates rapid analysis and is easily reduced (or enlarged) for rescan of specific genes of interest. The 20 nt sliding interval length is set such that sequences with the potential to form more than a single G4 DNA structure make correspondingly higher contributions to the total G4P. The last four windows analyzed are processed as windows of progressively smaller size (80, 60, 40 and 20 nt), and this does not affect the results as gene length is much larger than window size. Each DNA strand is evaluated independently. G4P is scored as a percentage, making it independent of sequence length. These criteria are similar to those used by others (4,5), and although not absolute, provide a means to compare G4P between different sequences. The software is written in C# to run on the Microsoft Windows XP operating system. The program and instructions are available on our laboratory website (http://depts.washington.edu/maizels9/). The source code is available upon request.

Sequence and GO data

Sequence data for the human RefSeq genes (NCBI 35 assembly) and associated GO terms were downloaded from the Ensembl database v.32 using BioMart (47) on 4/9/05. Additional Gene Ontology (GO) data were obtained from the GO website (www.geneontology.org) on 3/5/06. Flanking sequence data were downloaded from Ensembl v.34 on 14/10/05. The cDNA sequences were downloaded from the Ensembl v.37 on 19/3/06. For genes with multiple transcript variants, the first listed variant for each gene was evaluated to assess G4P of the cDNA. The median G4 DNA potential was calculated from 16 654 RefSeq genes. Since each gene may have several GO classifications, a total of 77 968 GO term assignments were sorted above or below the median corresponding to 50.5% or 49.5% of the total. The 4524 RefSeq genes that have no GO classification (27% of the total) are equally distributed below and above the median.

Flanking sequence analysis

Analysis of ΔG4P included only genes for which both gene and flanking sequence were complete. Excluded from that analysis were 101 genes for which sequence determination was incomplete (more than four unidentified consecutive bases). The excluded genes are identified in Supplementary Table 1; none of the excluded genes was a known proto-oncogene or a tumor suppressor gene.

Statistical analysis

The Wilcoxon rank sum test was applied by using the statistics program R 2.2.1 (Wilcoxon test parameters: alternative = 'two-sided', paired = FALSE). The linear regression, single-factor ANOVA, and standard error analyses were performed using Microsoft Office Excel 2003. Owing to the skewed distribution of G4P, data were subjected to a natural log transformation before linear regression analysis. Genes with G4P = 0 were therefore not included. This resulted in the exclusion of 1152 genes from the correlation between DNA strands (Figure 1B); and of 905 genes from the analysis of G4P versus GC content (Figure 3A).

CpG islands

The NewCpGReport software (48) was accessed and run from the website http://csc-fserve.hh.med.ic.ac.uk/emboss/newcpgreport.html. The tumor suppressor and proto-oncogene sequences were processed using the program's default settings: window size = 100; shift increment = 1; minimum length = 200; minimum observed/expected = 0.6; and minimum percent = 50.

RESULTS

G4 DNA formation potential of human genes

To score potential for G4 DNA formation, we developed software that analyzes overlapping windows of sequence, and scores each window that contains four or more runs of three or more guanines as a 'hit', then quantifies G4 DNA formation potential (G4P) as the percentage of hits in the total number of windows searched. We used this program, 'G4P Calculator', to evaluate the G4P of the entire transcribed sequence (exons and introns) of 16 654 human Reference Sequence (RefSeq) genes. Nontemplate and template strands were analyzed separately, to distinguish contributions

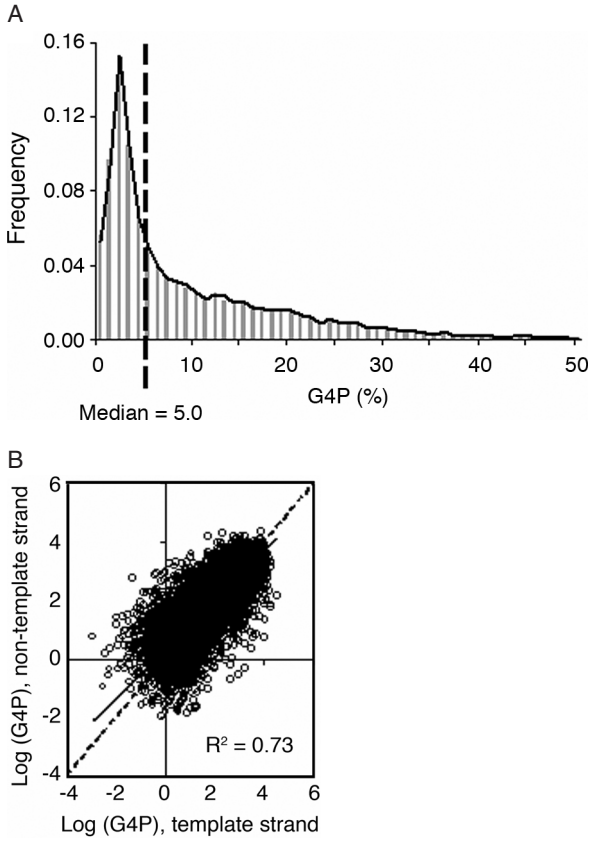

Figure 1. Potential for G4 DNA formation of human genes. (**A**) Distribution of genes across G4 DNA formation potential (G4P). The distribution of 16 654 RefSeq genes is illustrated by vertical bars (gray). Median G4P for the RefSeq genes at 5.0% is indicated by a dotted line. The distribution of the 4396 GO terms assigned to 73% of the RefSeq genes is outlined (black). (**B**) Positive correlation of G4P of template and nontemplate DNA strands. Linear regression analysis of G4P of the nontemplate (*y*-axis) and the template (*x*-axis) strand. Owing to the skewed distribution of G4P, the data were subjected to a natural log transformation before linear regression analysis; therefore, a small number of genes with G4P equal to zero were not included. The slope determined by linear regression analysis (0.83) is represented by the solid line; and slope of unity by the dotted line.

of transcription-induced structure formation, which affects only the nontemplate strand; and replication-induced structure formation, which affects both strands. G4P of the nontemplate strand ranged from 0 to 79%, with a median of 5.0% and an average of 9.1% (Figure 1A and Supplementary Table 1). Linear regression analysis (Figure 1B) showed that G4P of the nontemplate and template strands is positively correlated ($R^2 = 0.73$), with a slope of less than unity (0.83). Thus, for most genes, there is a slightly lower potential for the formation of transcription-induced structures than replication-induced structures; although potential for the formation of structures on either strand is closely correlated.

Gene function corresponds with G4P

The skewed distribution of G4P over the RefSeq genes (Figure 1A) suggests that most genes cannot readily form G4 DNA structures, but that some genes may be highly susceptible. To identify functional classes of genes with high and low potential for G4 DNA formation, we evaluated the distribution of terms defined by the GO Consortium (49) for each RefSeq gene across the spectrum of G4 DNA potentials. In this classification scheme, 27% of genes currently have no GO terms assigned; and others have been assigned multiple GO terms and are represented in several different categories. The distribution of the 4396 GO terms assigned to the human RefSeq genes across G4 DNA potential proved to be nearly identical to the distribution of genes (Figure 1A).

Restricting further analysis to the 218 GO terms associated with 50 or more genes, 61 GO terms were identified for which 60% or more of the genes were below or above the RefSeq median G4P (24 low G4P and 37 high G4P); and application of the Wilcoxon rank sum test confirmed that these criteria were robust (Table 1 and Supplementary Table 2). Functions characterized by low G4P include G-protein-coupled receptors, sensory perception (especially olfaction), nucleosome assembly, nucleic acid binding, ubiquitin cycle, cell adhesion and cell division; whereas functions characterized by high G4P include transcription factor activity, development, cell signaling, muscle contraction, growth factors and cytokines. Figure 2 represents, for a subset of the GO terms identified with very low and very high G4P, the median and range of G4P relative to all RefSeq genes and all GO terms. In each case, the difference in distribution relative to the RefSeq genes was highly significant (Figure 2). This establishes a relationship between specific gene functions and potential for G4 DNA formation.

Tumor suppressor genes are characterized by low G4P and proto-oncogenes by high G4P

Some of the gene functions characterized by low and high G4P (Figure 2 and Table 1) are associated with tumor suppressor genes and proto-oncogenes, respectively. This led us to interrogate the distribution of G4P with respect to genes in these two categories. A list of 55 tumor suppressor genes and 95 proto-oncogenes was compiled (Supplementary Table 3), using the Online Mendelian Inheritance in Man (OMIM) database as a primary source and confirming gene classification by search of the published literature. Comparison of G4P for tumor suppressor genes and proto-oncogenes established a clear and highly significant difference in the range of G4P observed (Wilcoxon rank sum test, $P = 10^{-8}$; Figure 3A), and in the distribution of G4P for genes in these two categories relative to the 16 654 genes in the RefSeq database (Figure 3B). The distribution of tumor suppressor genes was shifted from the RefSeq median of 5.0% towards low G4 DNA potential with a median of 2.4% (Wilcoxon rank sum test, $P = 4 \times 10^{-5}$); and the distribution of proto-oncogenes was shifted towards high G4P with a median of 11.0% (Wilcoxon rank sum test, $P = 7 \times 10^{-5}$).

Table 2 shows the top 10 genes in each category, ranked according to G4P, using a high stringency 40 nt search window. Analysis of G4P using a 40 nt rather than a 100 nt search window decreased the numerical value of G4P for each individual gene, as expected (Table 2 and Supplementary Table 3), but did not affect the relative differences in distribution of the potentials, and further supported the significance of the difference between tumor suppressor genes and proto-oncogenes (Wilcoxon rank sum test, $P = 2 \times 10^{-7}$). Table 2 also shows representative GO terms assigned to

Table 1. Gene Ontology (GO) terms with low and high G4P

	Biological process				Molecular function		
GO ID	GO description	No. of genes	*P*	GO ID	GO description	No. of genes	*P*
Low G4P							
GO:0007186	G-protein-coupled receptor protein signaling pathway	674	< 2E−16	GO:0004984	Olfactory receptor activity	316	< 2E−16
GO:0007600	Sensory perception	436	< 2E−16	GO:0003676	Nucleic acid binding	615	1E−06
GO:0007608	Perception of smell	244	< 2E−16	GO:0005488	Binding	427	5E−06
GO:0007001	Chromosome organization and biogenesis	99	8E−11	GO:0016874	Ligase activity	163	0.003
GO:0006334	Nucleosome assembly	95	9E−11	GO:0004197	Cysteine-type endopeptidase activity	55	0.006
GO:0006511	Ubiquitin-dependent protein catabolism	96	5E−05	GO:0004842	Ubiquitin–protein ligase activity	327	0.02
GO:0006512	Ubiquitin cycle	210	0.0002	GO:0017111	Nucleoside triphosphatase activity	54	0.03
GO:0007156	Homophilic cell adhesion	87	0.02	GO:0008026	ATP-dependent helicase activity	63	0.03
GO:0051301	Cell division	119	0.03				
GO:0006470	Protein amino acid dephosphorylation	115	0.08				
High G4P							
GO:0007275	Development	391	< 2E−16	GO:0003700	Transcription factor activity	752	< 2E−16
GO:0006955	Immune response	272	3E−09	GO:0004295	Trypsin activity	94	2E−07
GO:0007267	Cell–cell signaling	263	2E−06	GO:0004263	Chymotrypsin activity	91	9E−07
GO:0006936	Muscle contraction	63	6E−05	GO:0030528	Transcription regulator activity	74	8E−05
GO:0006817	Phosphate transport	80	7E−05	GO:0030955	Potassium ion binding	100	0.0005
GO:0007010	Cytoskeleton organization and biogenesis	53	0.0002	GO:0008083	Growth factor activity	116	0.003
GO:0009653	Morphogenesis	104	0.002	GO:0005179	Hormone activity	80	0.004
GO:0007517	Muscle development	103	0.002	GO:0008289	Lipid binding	75	0.008
GO:0007218	Neuropeptide signaling pathway	66	0.01	GO:0003774	Motor activity	58	0.02
GO:0006814	Sodium ion transport	85	0.02	GO:0015293	Symporter activity	71	0.02
GO:0006968	Cellular defense response	61	0.02	GO:0005249	Voltage-gated potassium channel activity	65	0.02
GO:0006954	Inflammatory response	163	0.02	GO:0005125	Cytokine activity	79	0.05
GO:0006091	Generation of precursor metabolites and energy	59	0.02	GO:0020037	Heme binding	78	0.06
GO:0001501	Skeletal development	78	0.03				
GO:0007169	Transmembrane receptor protein tyrosine kinase signaling pathway	67	0.04				
GO:0008544	Epidermis development	62	0.04				
GO:0006816	Calcium ion transport	63	0.08				
GO:0006869	Lipid transport	53	0.08				
GO:0009887	Organogenesis	50	0.09				

GO term ID number, description and number of genes to which this term applies, for each GO term associated with Biological Processes and Molecular Functions, and containing genes with a distribution that is significantly lower or higher in G4P than the RefSeq genes. Terms are sorted by ascending *P*-value (shown) as calculated by the Wilcoxon rank sum test. Additional data can be found in Supplementary Table 2.

each gene, many of which correspond to GO terms overrepresented in low or high G4P (Figure 2).

Simulations were carried out to verify the significance of the differences between G4P of tumor suppressor genes and proto-oncogenes. G4P distributions of 95 or 55 genes picked at random were not significantly different from the RefSeq set ($P > 0.1$), in each of 20 iterations. Furthermore, statistical significance of observed differences was robust to misclassification of up to 10 genes per category, as tested by the addition of 10 randomly selected RefSeq genes to either category ($P < 0.002$), or by elimination of 10 randomly selected genes from either category ($P < 0.0002$). These simulations confirmed the significance of the differences between G4P of tumor suppressor genes and proto-oncogenes.

Tumor suppressor genes and proto-oncogenes have similar numbers of CpG islands

CpG islands are associated with a majority of promoters of human genes (50), and CpG dinucleotides are targets for

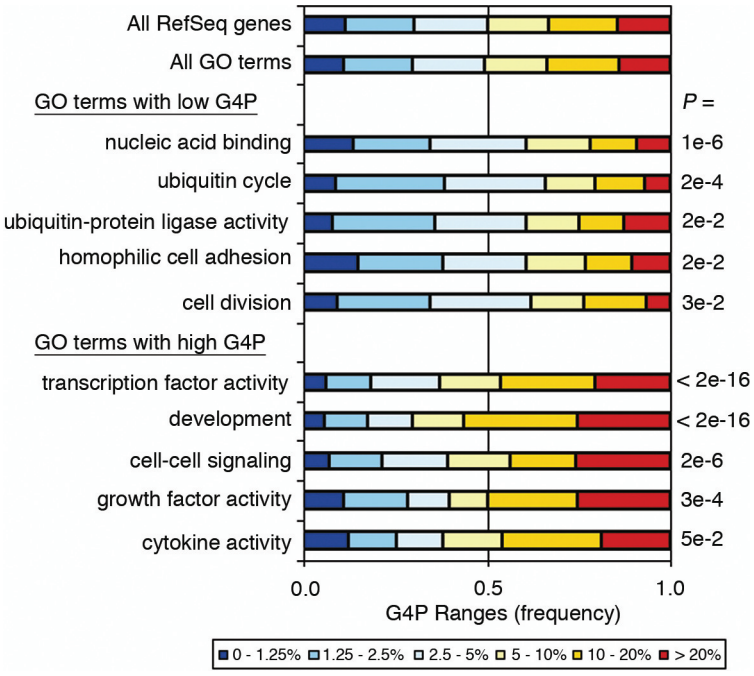

Figure 2. G4P correlates with gene function. Ranges of G4P for all RefSeq genes (top line) compared with five GO terms overrepresented in low or high G4P. Boxes represent the percentage of genes in each GO category characterized by G4P in the range 0–1.25, 1.25–2.5, 2.5–5.0, 5–10, 10–20% and >20% (colors as indicated). *P*-values shown on the right represent significance of the difference in distribution between each GO term and the RefSeq genes, as calculated by the Wilcoxon rank sum test.

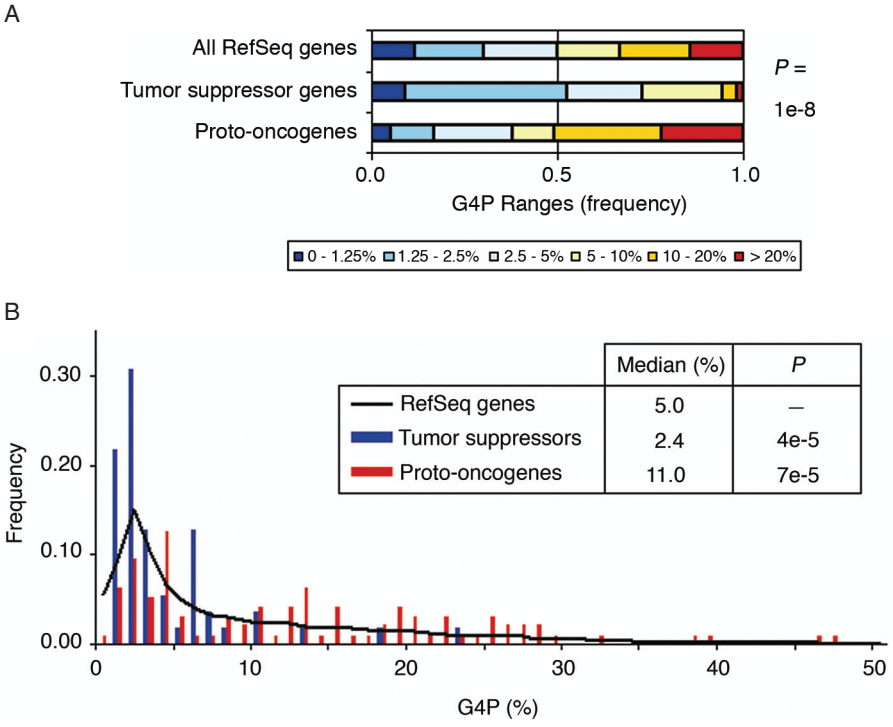

Figure 3. Contrasting G4P of tumor suppressor genes and proto-oncogenes. (**A**) Ranges of G4P for 55 tumor suppressor genes, 95 proto-oncogenes and all 16 654 RefSeq genes. Boxes represent the percentage of genes in each category characterized by G4P in the range 0–1.25, 1.25–2.5, 2.5–5.0, 5–10, 10–20% and >20% (colors as indicated). *P*-value represents significance of the difference in distribution between the tumor suppressor genes and proto-oncogenes, as calculated by the Wilcoxon rank sum test. (**B**) Distribution of tumor suppressor genes and proto-oncogenes across G4P. Bars represent the G4P distribution of 55 tumor suppressor genes (blue) and 95 proto-oncogenes (red). The black outline diagrams distribution of all 16 654 RefSeq genes (as in Figure 1A). *P*-values represent significance of the difference in distribution between each group of genes and the RefSeq genes, as calculated by the Wilcoxon rank sum test.

Table 2. Tumor suppressor genes with low G4P and proto-oncogenes with high G4P

Tumor suppressor genes				Proto-oncogenes			
HGNC symbol	G4P, 100 nt window (%)	G4P, 40 nt window (%)	Representative GO terms	HGNC symbol	G4P, 100 nt window (%)	G4P, 40 nt window (%)	Representative GO terms
FBXW7	0.5	0.00	Protein ubiquitination	*FGF4*	46.6	6.8	Cell–cell signaling, growth factor activity
MAD2L1	0.8	0.00	Cell cycle, cell division	*AKT1*	45.9	5.3	Anti-apoptosis, signal transduction
SMARCA3	0.8	0.00	Ubiquitin–protein ligase activity, DNA binding	*HRAS*	37.6	4.3	Organogenesis, GTPase activity
APC	1.1	0.00	Cell adhesion, negative regulation of cell cycle	*IGF2*	31.6	4.1	Development, growth factor activity, hormone activity
BLM	1.4	0.00	DNA binding, DNA repair	*BCL3*	25.4	4.1	Transcription, regulation of cell cycle
THBS1	1.5	0.00	Cell adhesion, cell motility	*NOTCH1*	39.2	3.9	Transcription factor activity, epidermis development
VHL	2.7	0.00	Protein ubiquitination, negative regulation of cell cycle	*NFKB2*	27.6	3.5	Transcription factor activity, signal transduction
CDKN2B	6.6	0.00	Negative regulation of cell cycle	*FURIN*	27.8	3.4	Cell–cell signaling
MLL3	0.9	0.02	Ubiquitin–protein ligase activity, DNA binding	*JUNB*	25.6	3.4	Transcription factor activity
BRCA2	1.5	0.05	Nucleic acid binding, DNA repair, regulation of cell cycle	*GLI1*	20.3	2.8	Development, transcription, signal transduction

The table lists the top 10 genes in each group, sorted by G4P (40 nt search window). The HGNC symbol and values for G4P (both 40 and 100 nt search windows) are shown, along with representative GO terms for each of the genes. The complete list of 55 tumor suppressor genes and 95 proto-oncogenes is available as Supplementary Table 3.

methylation leading to gene silencing [reviewed in (51)]. G4P is positively correlated with GC-content (Figure 4A), which could in principle reflect a local enrichment of CpG methylation sites. We tested this possibility by analyzing tumor suppressor genes and proto-oncogenes with the EMBOSS program 'NewCpGReport' (48), which identifies CpG-rich regions. The number of CpG-rich regions do not differ significantly between these two categories of genes (Wilcoxon rank sum test, $P = 0.4$; Figure 4B). Thus, the density of potential methylation sites does not distinguish tumor suppressor genes from proto-oncogenes, or account for the differences in G4P we have documented.

Both exon and intron sequences contribute to G4P

In human genes, the ratio of exon length to intron length is typically well below unity (52), so the measurement of G4P for an entire gene will largely reflect the contribution of intronic sequences. To distinguish contributions of exons and introns to G4P, we analyzed the G4P of one representative cDNA for each tumor suppressor gene and proto-oncogene (Figure 5 and Supplementary Table 3). The median G4P of cDNA sequences is 1.9% for tumor suppressor genes, and 7.6% for proto-oncogenes (Figure 5), in each case slightly lower than G4P for the entire gene (2.4 and 11%, respectively; Figure 3B). The difference between G4P for cDNAs in the two functional categories is highly significant (Wilcoxon rank sum test, $P = 6 \times 10^{-6}$). Thus, both exon and intron sequences contribute to the significant differences in G4P characteristic of tumor suppressor and proto-oncogenes.

The G4P of genes contrasts with their genomic environment

The human genome consists of large segments of fairly homogeneous GC-content, defined as isochores (53). With the availability of the human genome sequence, this definition has been honed further, and 100 kb segments of DNA sequence can be sorted into five isochore families with an average SD of ~1% GC (54). Since G4P is positively correlated with GC-content (Figure 4A), we asked if G4P for each gene reflects its local genomic environment. To do this, we computed the difference in G4P for each of the RefSeq genes and its flanking sequences, $\Delta G4P = G4P - G4P_{FLANK}$, calculating $G4P_{FLANK}$ as the average G4P for 20 kb upstream and downstream of each gene (Supplementary Table 1 and Figure 6). The average $\Delta G4P$ for all RefSeq genes is 1.6%; thus on average, genes have greater G4P than their flanking sequences. Comparison of $\Delta G4P$ of the RefSeq genes, the proto-oncogenes and the tumor suppressor genes by a single-factor ANOVA showed that the three groups are distinct (ANOVA, $P = 10^{-5}$; Figure 6). The average $\Delta G4P$ for the set of tumor suppressor genes is −2.2%, much lower than that of the RefSeq genes (ANOVA, $P = 5 \times 10^{-5}$). In contrast, the average $\Delta G4P$ for the set of proto-oncogenes is 3.4%, higher than that of the RefSeq genes (ANOVA, $P = 0.01$), and considerably higher than that of the tumor suppressor genes (ANOVA, $P = 5 \times 10^{-7}$). Thus, on average, tumor suppressor genes have lower G4P than their flanking sequences, and proto-oncogenes have higher G4P than their flanking sequences. Potential for G4 DNA formation therefore correlates with

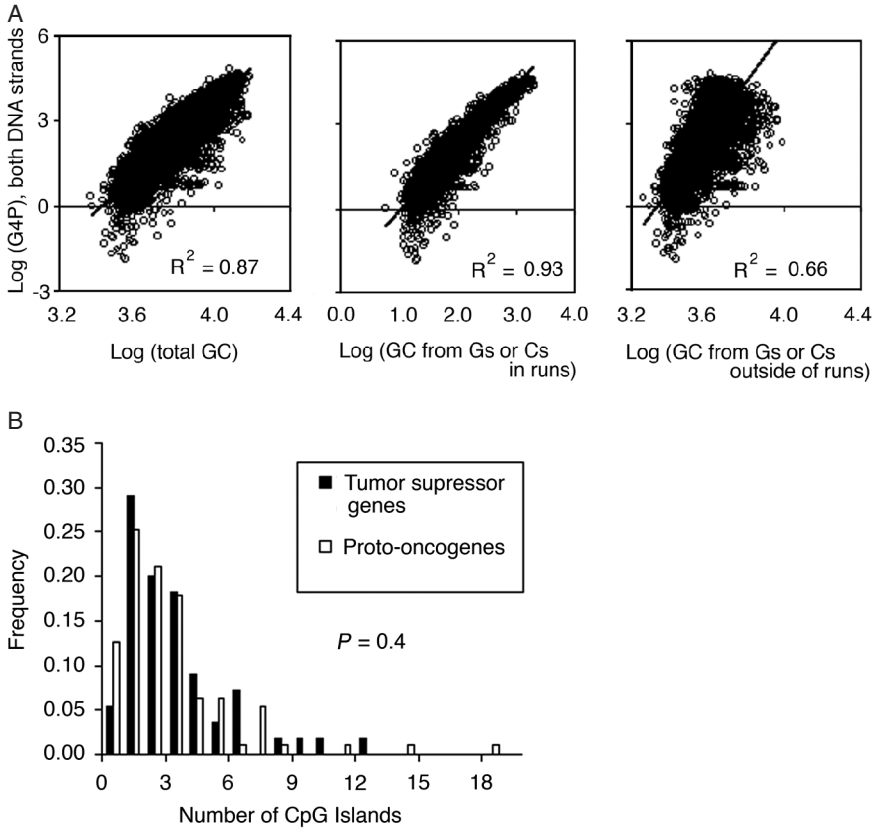

Figure 4. G4P correlates with GC-content but not CpG islands. (**A**) Correlation of G4P with GC-content. Linear regression analysis of G4P relative to total GC-content (left); the portion of GC-content contributed from G-runs or C-runs (center); the remaining GC-content contributed from Gs and Cs outside of G-runs or C-runs (right). The data were subjected to a natural log transformation before linear regression analysis; therefore, a small number of genes with G4P equal to zero were not included. The slopes determined by linear regression analysis are represented by solid lines. G4P correlates most closely with Gs and Cs within runs (middle). (**B**) Distribution of tumor suppressor genes and proto-oncogenes relative to number of CpG islands. Closed bars, tumor suppressor genes; open bars, proto-oncogenes. *P*-value was determined by the Wilcoxon rank sum test comparing tumor suppressor genes to proto-oncogenes, and shows that there is not a significant relationship between gene function and number of CpG islands.

gene function, rather than local genomic environment, for both tumor suppressor genes and proto-oncogenes.

The prototypical tumor suppressor and proto-oncogenes, *TP53* and *MYC* [reviewed in (55)], illustrate the relationship between G4P of genes and their flanking sequences. For *TP53*, G4P is 7.6%, slightly higher than the RefSeq median of 5%; and $G4P_{FLANK}$ is 12.1% (10.0% upstream and 14.1% downstream). Thus, ΔG4P of *TP53* is −4.4%, low even among tumor suppressor genes. In contrast, for *MYC*, G4P is 18.6%, well above the RefSeq median; and $G4P_{FLANK}$ is 2.8% (3.3% upstream and 2.2% downstream). Thus, ΔG4P of *MYC* is 15.9%, considerably above the average RefSeq ΔG4P of 1.6%.

DISCUSSION

We have investigated the relationship between potential to form G4 DNA and gene function for the 16 654 human RefSeq genes. We find that there is a highly skewed distribution of G4P among human genes, and that there are robust correlations between G4P and gene function. Interrogation of the subset of 218 GO terms assigned to 50 or more genes showed that low G4P corresponds with functions including G-protein-coupled receptors, olfaction, nucleosome assembly, nucleic acid binding, ubiquitin cycle, cell adhesion and cell division; and high G4P with functions including transcription factor activity, development, cell signaling, growth factors and cytokines. These findings motivated interrogation of two contrasting gene categories defined by the OMIM database, tumor suppressor genes and proto-oncogenes, which showed that genes in these categories are distinguished by low and high G4P, respectively (Figure 3).

In contrast to the robust relationship between G4P and gene function, G4P did not correspond to any of several well-established parameters used to characterize genomic structure. G4P does correlate with GC-content, but not with the number of CpG islands (Figure 4). Both exons and introns contribute to the difference in G4P between tumor suppressor genes and proto-oncogenes (Figure 5). Furthermore, G4P does not reflect the local genomic environment (Figure 6). In fact, tumor suppressor genes have much lower G4P than would be predicted by their genomic environment as compared to the RefSeq genes, whereas proto-oncogenes have higher G4P than would be predicted. The most straightforward interpretation of these results is that genes with specific functions have undergone selection based on G4P.

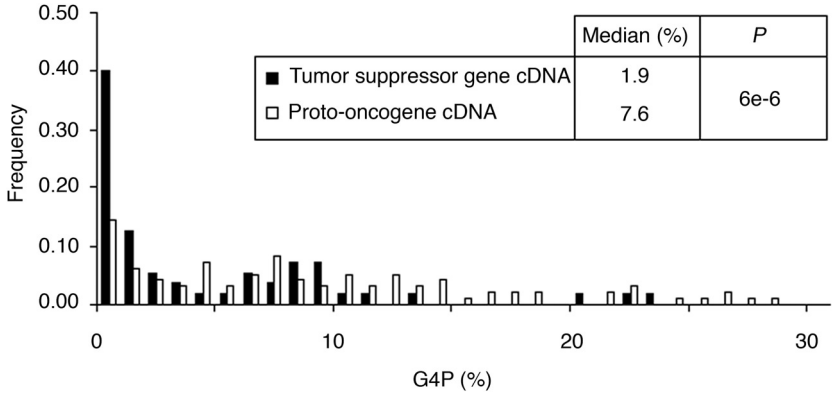

Figure 5. Differences in G4P of tumor suppressor and proto-oncogene cDNAs. Distribution of tumor suppressor gene and proto-oncogene cDNA sequences across G4P. Bars represent tumor suppressor genes (closed bars) and proto-oncogenes (open bars). *P*-value was determined from the Wilcoxon rank sum test comparing tumor suppressor genes to proto-oncogenes.

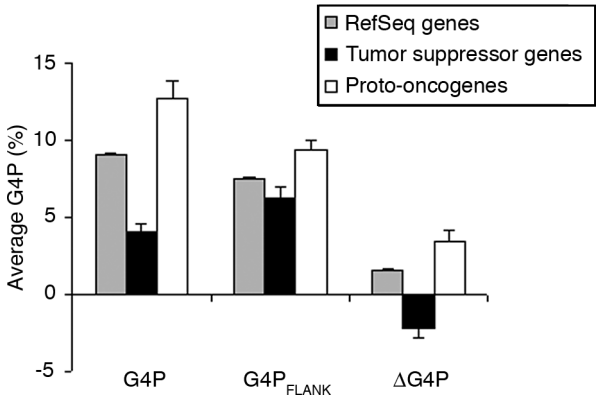

Figure 6. G4P of genes differs from G4P of genomic environment. Average G4P for genes; 20 kb flanking sequences ($G4P_{FLANK}$); and ΔG4P, the difference between G4P for each gene and its flank. Gray bars, RefSeq genes; closed bars, tumor suppressor genes; and open bars, proto-oncogenes. Standard errors were determined by ANOVA for each analysis of the three groups of genes.

One source of selective pressure that could contribute to determining G4P is suggested by the association between G-rich regions and genomic instability. Transcription-induced or replication-induced DNA structures can form within regions of high G4P, and if these structures are not faithfully resolved, the result may be genomic instability and impaired gene function. In this view, the low G4P of tumor suppressor genes could reflect evolution that minimized potential instability of genes which function to maintain genomic stability. There is considerable evidence for haploinsufficiency of tumor suppressor genes [reviewed in (56)], and this would contribute to pressure to minimize genomic instability. Conversely, the high G4P that characterizes the proto-oncogenes would be predicted to contribute to their destabilization. Could *instability* provide a selective advantage? Under some circumstances, it may. Proto-oncogenes are transcribed in rapidly dividing cells and tissues. Transcription-induced structures have considerable potential to contribute to genomic instability (25,26), but they can form only within genes, which represent a relatively small fraction of genomic DNA. The high G4P of the proto-oncogenes would make them targets for transcription-induced destabilization. Proto-oncogenes encode key factors that promote cell proliferation and development, and impaired expression of a proto-oncogene could in turn diminish or prevent cell proliferation, either by decreasing expression of an essential factor, or signaling cell death via apoptosis. Proto-oncogenes may therefore carry out a passive surveillance function, monitoring instability that specifically affected the transcribed fraction of the genome. This surveillance function would necessarily be vested in genes, rather than in the vast landscape of nontranscribed sequences, consistent with the clear differences between G4P of genes and their flanking sequences.

Another mechanism that may contribute to selection based on G4P is shared regulation. Sequences within promoter regions of several proto-oncogenes have been shown to form G4 DNA *in vitro* (29–32), and factors that bind G4 DNA have been implicated in both transcriptional and translational regulation (28,57). However, regulatory factors typically exert their effects within limited genomic regions, so commonality of short *cis*-regulatory elements is unlikely to provide a complete explanation for a feature of sequence composition that distinguishes both exons and introns, and extends throughout a gene (Figure 5). Similarly, G4P is unlikely to reflect selection for coding capacity, as this sort of selection would affect exons alone. Nonetheless, there does appear to be some selection against regions of high G4P within exons, as in both gene categories, the median G4P of exons was lower than for introns: 1.9% versus 2.4% for tumor suppressor genes; and 7.6% versus 11% for proto-oncogenes (Figures 3B and 5). Thus high G4P may be disfavored in mature RNAs, as has been proposed previously (4); or incompatible with efficient translation or effective coding.

Several lines of evidence suggest that GC-content may broadly correlate with gene expression levels (58–60); in particular, GC-richness correlates with open chromatin structure, which may in turn facilitate transcription (61). Proto-oncogenes are rapidly transcribed during early development and in response to cell activation, and the high G4P of the proto-oncogenes might reflect GC-richness that contributes to high transcription levels of genes in this group. The finding that potential for G4 DNA formation correlates robustly with specific gene functions suggests that G4P may be a useful

parameter to include in global analyses of gene expression, regulation and interactions. Systems-based analyses of this sort should establish whether regulation could contribute to selection based on G4P.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank Audrey Qiuyan Fu and Paul Sampson for assistance with statistical analysis; John Newman for independent testing of the G4P Calculator software; and Evan Eichler for comments on the manuscript. This work was supported by NIH R01 GM65988 to N.M. and J.E. was supported by T32 CA009537. Funding to pay the Open Access publication charges for this article was provided by NIH RO1 GM65988.

Conflict of interest statement. None declared.

REFERENCES

1. Sen,D. and Gilbert,W. (1988) Formation of parallel four-stranded complexes by guanine-rich motifs in DNA and its implications for meiosis. *Nature*, **334**, 364–366.
2. Gellert,M., Lipsett,M.N. and Davies,D.R. (1962) Helix formation by guanylic acid. *Proc. Natl Acad. Sci. USA*, **48**, 2013–2018.
3. Phan,A.T., Kuryavyi,V. and Patel,D.J. (2006) DNA architecture: from G to Z. *Curr. Opin. Struct. Biol.*, **16**, 288–298.
4. Hazel,P., Parkinson,G.N. and Neidle,S. (2006) Predictive modelling of topology and loop variations in dimeric DNA quadruplex structures. *Nucleic Acids Res.*, **34**, 2117–2127.
5. Huppert,J.L. and Balasubramanian,S. (2005) Prevalence of quadruplexes in the human genome. *Nucleic Acids Res.*, **33**, 2908–2916.
6. Todd,A.K., Johnston,M. and Neidle,S. (2005) Highly prevalent putative quadruplex sequence motifs in human DNA. *Nucleic Acids Res.*, **33**, 2901–2907.
7. Zahler,A.M., Williamson,J.R., Cech,T.R. and Prescott,D.M. (1991) Inhibition of telomerase by G-quartet DNA structures. *Nature*, **350**, 718–720.
8. Ishikawa,F., Matunis,M.J., Dreyfuss,G. and Cech,T.R. (1993) Nuclear proteins that bind the pre-mRNA 3′ splice site sequence r(UUAG/G) and the human telomeric DNA sequence d(TTAGGG)$_n$. *Mol. Cell. Biol.*, **13**, 4301–4310.
9. Fletcher,T.M., Sun,D., Salazar,M. and Hurley,L.H. (1998) Effect of DNA secondary structure on human telomerase activity. *Biochemistry*, **37**, 5536–5541.
10. LaBranche,H., Dupuis,S., Ben-David,Y., Bani,M.-R., Wellinger,R.J. and Chabot,B. (1998) Telomere elongation by hnRNP A1 and a derivative that interacts with telomeric repeats and telomerase. *Nature Genet.*, **19**, 1–4.
11. Eversole,A. and Maizels,N. (2000) *In vitro* properties of the conserved mammalian protein hnRNP D suggest a role in telomere maintenance. *Mol. Cell. Biol.*, **20**, 5425–5432.
12. Enokizono,Y., Konishi,Y., Nagata,K., Ouhashi,K., Uesugi,S., Ishikawa,F. and Katahira,M. (2005) Structure of hnRNP D complexed with single-stranded telomere DNA and unfolding of the quadruplex by heterogeneous nuclear ribonucleoprotein D. *J. Biol. Chem.*, **280**, 18862–18870.
13. Paeschke,K., Simonsson,T., Postberg,J., Rhodes,D. and Lipps,H.J. (2005) Telomere end-binding proteins control the formation of G-quadruplex DNA structures *in vivo*. *Nature Struct. Mol. Biol.*, **12**, 847–854.
14. Zaug,A.J., Podell,E.R. and Cech,T.R. (2005) Human POT1 disrupts telomeric G-quadruplexes allowing telomerase extension *in vitro*. *Proc. Natl Acad. Sci. USA*, **102**, 10864–10869.
15. Zhang,Q.S., Manche,L., Xu,R.M. and Krainer,A.R. (2006) hnRNP A1 associates with telomere ends and stimulates telomerase activity. *RNA*, **12**, 1116–1128.
16. Maizels,N. (2005) Immunoglobulin gene diversification. *Annu. Rev. Genet.*, **39**, 23–46.
17. Larson,E.D., Duquette,M.L., Cummings,W.J., Streiff,R.J. and Maizels,N. (2005) MutSalpha binds to and promotes synapsis of transcriptionally activated immunoglobulin switch regions. *Curr. Biol.*, **15**, 470–474.
18. Pasqualucci,L., Neumeister,P., Goossens,T., Nanjangud,G., Chaganti,R.S., Kuppers,R. and Dalla-Favera,R. (2001) Hypermutation of multiple proto-oncogenes in B-cell diffuse large-cell lymphomas. *Nature*, **412**, 341–346.
19. 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.
20. Duquette,M.L., Pham,P., Goodman,M.F. and Maizels,N. (2005) AID binds to transcription-induced structures in c-MYC that map to regions associated with translocation and hypermutation. *Oncogene*, **24**, 5791–5798.
21. Wong,Z., Wilson,V., Patel,I., Povey,S. and Jeffreys,A.J. (1987) Characterization of a panel of highly variable minisatellites cloned from human DNA. *Ann. Hum. Genet.*, **51**, 269–288.
22. Weitzmann,M.N., Woodford,K.J. and Usdin,K. (1997) DNA secondary structures and the evolution of hypervariable tandem arrays. *J. Biol. Chem.*, **272**, 9517–9523.
23. Kilburn,A.E., Shea,M.J., Sargent,R.G. and Wilson,J.H. (2001) Insertion of a telomere repeat sequence into a mammalian gene causes chromosome instability. *Mol. Cell. Biol.*, **21**, 126–135.
24. Reaban,M.E. and Griffin,J.A. (1990) Induction of RNA-stabilized DNA conformers by transcription of an immunoglobulin switch region. *Nature*, **348**, 342–344.
25. Mizuta,R., Iwai,K., Shigeno,M., Mizuta,M., Ushiki,T. and Kitamura,D. (2002) Molecular visualization of immunoglobulin switch region RNA/DNA complex by atomic force microscope. *J. Biol. Chem.*, **278**, 4431–4434.
26. Huertas,P. and Aguilera,A. (2003) Cotranscriptionally formed DNA:RNA hybrids mediate transcription elongation impairment and transcription-associated recombination. *Mol. Cell*, **12**, 711–721.
27. Li,X. and Manley,J.L. (2005) Inactivation of the SR protein splicing factor ASF/SF2 results in genomic instability. *Cell*, **122**, 365–378.
28. Darnell,J.C., Fraser,C.E., Mostovetsky,O., Stefani,G., Jones,T.A., Eddy,S.R. and Darnell,R.B. (2005) Kissing complex RNAs mediate interaction between the Fragile-X mental retardation protein KH2 domain and brain polyribosomes. *Genes Dev.*, **19**, 903–918.
29. Darnell,J.C., Jensen,K.B., Jin,P., Brown,V., Warren,S.T. and Darnell,R.B. (2001) Fragile X mental retardation protein targets G quartet mRNAs important for neuronal function. *Cell*, **107**, 489–499.
30. Simonsson,T., Pecinka,P. and Kubista,M. (1998) DNA tetraplex formation in the control region of c-myc. *Nucleic Acids Res.*, **26**, 1167–1172.
31. 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.
32. 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.
33. Dai,J., Dexheimer,T.S., Chen,D., Carver,M., Ambrus,A., Jones,R.A. and Yang,D. (2006) An intramolecular G-quadruplex structure with mixed parallel/antiparallel G-strands formed in the human BCL-2 promoter region in solution. *J. Am. Chem. Soc.*, **128**, 1096–1098.
34. Rezler,E.M., Bearss,D.J. and Hurley,L.H. (2003) Telomere inhibition and telomere disruption as processes for drug targeting. *Annu. Rev. Pharmacol. Toxicol.*, **43**, 359–379.
35. Gomez,D., Lemarteleur,T., Lacroix,L., Mailliet,P., Mergny,J.L. and Riou,J.F. (2004) Telomerase downregulation induced by the G-quadruplex ligand 12459 in A549 cells is mediated by hTERT RNA alternative splicing. *Nucleic Acids Res.*, **32**, 371–379.

36. Burger,A.M., Dai,F., Schultes,C.M., Reszka,A.P., Moore,M.J., Double,J.A. and Neidle,S. (2005) The G-quadruplex-interactive molecule BRACO-19 inhibits tumor growth, consistent with telomere targeting and interference with telomerase function. *Cancer Res*., **65**, 1489–1496.
37. Tauchi,T., Shin-Ya,K., Sashida,G., Sumi,M., Okabe,S., Ohyashiki,J.H. and Ohyashiki,K. (2006) Telomerase inhibition with a novel G-quadruplex-interactive agent, telomestatin: *in vitro* and *in vivo* studies in acute leukemia. *Oncogene* (in press).
38. 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.
39. 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.
40. Huber,M.D., Lee,D.C. and Maizels,N. (2002) G4 DNA unwinding by BLM and Sgs1p: substrate specificity and substrate-specific inhibition. *Nucleic Acids Res*., **30**, 3954–3961.
41. Fry,M. and Loeb,L.A. (1999) Human werner syndrome DNA helicase unwinds tetrahelical structures of the fragile X syndrome repeat sequence d(CGG)n. *J. Biol. Chem*., **274**, 12797–12802.
42. Sun,H., Bennett,R.J. and Maizels,N. (1999) The *S. cerevisiae* Sgs1 helicase efficiently unwinds G–G paired DNAs. *Nucleic Acids Res*., **27**, 1978–1984.
43. Sinclair,D.A. and Guarente,L. (1997) Extrachromosomal rDNA circles—a cause of aging in yeast. *Cell*, **91**, 1033–1042.
44. Versini,G., Comet,I., Wu,M., Hoopes,L., Schwob,E. and Pasero,P. (2003) The yeast Sgs1 helicase is differentially required for genomic and ribosomal DNA replication. *EMBO J*., **22**, 1939–1949.
45. Crabbe,L., Verdun,R.E., Haggblom,C.I. and Karlseder,J. (2004) Defective telomere lagging strand synthesis in cells lacking WRN helicase activity. *Science*, **306**, 1951–1953.
46. Yang,Q., Zhang,R., Wang,X.W., Linke,S.P., Sengupta,S., Hickson,I.D., Pedrazzi,G., Perrera,C., Stagljar,I., Littman,S.J. *et al*. (2004) The mismatch DNA repair heterodimer, hMSH2/6, regulates BLM helicase. *Oncogene*, **23**, 3749–3756.
47. Hubbard,T., Andrews,D., Caccamo,M., Cameron,G., Chen,Y., Clamp,M., Clarke,L., Coates,G., Cox,T., Cunningham,F. *et al*. (2005) Ensembl 2005. *Nucleic Acids Res*., **33**, D447–D453.
48. Rice,P., Longden,I. and Bleasby,A. (2000) EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet*., **16**, 276–277.
49. Ashburner,M., Ball,C.A., Blake,J.A., Botstein,D., Butler,H., Cherry,J.M., Davis,A.P., Dolinski,K., Dwight,S.S., Eppig,J.T. *et al*. (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature Genet*., **25**, 25–29.
50. Carninci,P., Sandelin,A., Lenhard,B., Katayama,S., Shimokawa,K., Ponjavic,J., Semple,C.A., Taylor,M.S., Engstrom,P.G., Frith,M.C. *et al*. (2006) Genome-wide analysis of mammalian promoter architecture and evolution. *Nature Genet*., **38**, 626–635.
51. Klose,R.J. and Bird,A.P. (2006) Genomic DNA methylation: the mark and its mediators. *Trends Biochem. Sci*., **31**, 89–97.
52. Lander,E.S., Linton,L.M., Birren,B., Nusbaum,C., Zody,M.C., Baldwin,J., Devon,K., Dewar,K., Doyle,M., FitzHugh,W. *et al*. (2001) Initial sequencing and analysis of the human genome. *Nature*, **409**, 860–921.
53. Bernardi,G., Olofsson,B., Filipski,J., Zerial,M., Salinas,J., Cuny,G., Meunier-Rotival,M. and Rodier,F. (1985) The mosaic genome of warm-blooded vertebrates. *Science*, **228**, 953–958.
54. Costantini,M., Clay,O., Auletta,F. and Bernardi,G. (2006) An isochore map of human chromosomes. *Genome Res*., **16**, 536–541.
55. Hanahan,D. and Weinberg,R.A. (2000) The hallmarks of cancer. *Cell*, **100**, 57–70.
56. Payne,S.R. and Kemp,C.J. (2005) Tumor suppressor genetics. *Carcinogenesis*, **26**, 2031–2045.
57. 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.
58. Versteeg,R., van Schaik,B.D., van Batenburg,M.F., Roos,M., Monajemi,R., Caron,H., Bussemaker,H.J. and van Kampen,A.H. (2003) The human transcriptome map reveals extremes in gene density, intron length, GC content, and repeat pattern for domains of highly and weakly expressed genes. *Genome Res*., **13**, 1998–2004.
59. Lercher,M.J., Urrutia,A.O., Pavlicek,A. and Hurst,L.D. (2003) A unification of mosaic structures in the human genome. *Hum. Mol. Genet*., **12**, 2411–2415.
60. Semon,M., Mouchiroud,D. and Duret,L. (2005) Relationship between gene expression and GC-content in mammals: statistical significance and biological relevance. *Hum. Mol. Genet*., **14**, 421–427.
61. Gilbert,N., Boyle,S., Fiegler,H., Woodfine,K., Carter,N.P. and Bickmore,W.A. (2004) Chromatin architecture of the human genome: gene-rich domains are enriched in open chromatin fibers. *Cell*, **118**, 555–566.