Published online 10 January 2008 *Nucleic Acids Research, 2008, Vol. 36, No. 4* ***1321–1333***
doi:10.1093/nar/gkm1138

Conserved elements with potential to form polymorphic G-quadruplex structures in the first intron of human genes

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

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

Received November 5, 2007; Revised December 5, 2007; Accepted December 6, 2007

ABSTRACT

To understand how potential for G-quadruplex formation might influence regulation of gene expression, we examined the 2 kb spanning the transcription start sites (TSS) of the 18 217 human RefSeq genes, distinguishing contributions of template and nontemplate strands. Regions both upstream and downstream of the TSS are G-rich, but the downstream region displays a clear bias toward G-richness on the nontemplate strand. Upstream of the TSS, much of the G-richness and potential for G-quadruplex formation derives from the presence of well-defined canonical regulatory motifs in duplex DNA, including CpG dinucleotides which are sites of regulatory methylation, and motifs recognized by the transcription factor SP1. This challenges the notion that quadruplex formation upstream of the TSS contributes to regulation of gene expression. Downstream of the TSS, G-richness is concentrated in the first intron, and on the nontemplate strand, where polymorphic sequence elements with potential to form G-quadruplex structures and which cannot be accounted for by known regulatory motifs are found in almost 3000 (16%) of the human RefSeq genes, and are conserved through frogs. These elements could in principle be recognized either as DNA or as RNA, providing structural targets for regulation at the level of transcription or RNA processing.

INTRODUCTION

Genomic DNA sequences are not random, but relatively little is known about how their non-randomness contributes to biological functions such as gene expression, genome stability and evolution. Among the most intriguing non-random sequences are G-rich regions, which characterize single-copy genes and also repetitive genomic domains including telomeres, ribosomal DNA and the immunoglobulin heavy-chain switch regions. G-rich DNA sequences have the potential to form G4 DNA (also known as G-quadruplex DNA or G-tetraplex DNA), a structure in which intra- or inter-strand interactions are stabilized by G-quartets, planar arrays of four guanines, paired by Hoogsteen bonding (1–3). G4 DNA forms spontaneously in synthetic oligonucleotides which contain at least four runs of guanines, with at least three guanines per run. RNA can form a similar structure, and G4 DNA and G4 RNA are both very stable once formed, with stability derived from hydrogen bonding between guanines and stacking of G-quartets, as well as by the length of the guanine runs and the intervening sequences that form the loops of the structures (4,5). Systematic analysis of quadruplexes formed by synthetic oligonucleotides has shown that there is an enormous potential for structural diversity, and that the strands that connect the stacked quartets may be antiparallel, parallel or a mix of these orientations (1). Many genomic G-rich regions carry more than four G-runs, creating the potential for combinatorial diversity that could contribute to formation of polymorphic G-quadruplex structures.

For G4 DNA to form in the genome of a living cell, G-rich regions must be released from the DNA duplex, as occurs during transient denaturation that accompanies replication, transcription and recombination (2). G4 DNA is recognized by a number of conserved proteins, including RecQ family helicases and MutSα (6–10), which may remove G4 DNA formed during replication or transcription to maintain genomic stability; nucleolin (11), the major component of the vertebrate nucleolus, where the G-rich rDNA is transcribed and rRNA biogenesis occurs; and factors associated with mRNA processing, including hnRNP D and hnRNP A1 (12,13). Some proteins that interact with G4 DNA do so with high affinity (nanomolar), consistent with the notion that

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

G-quadruplexes could provide a target for regulation of genomic stability or gene expression *in vivo*.

In the human genome, the number of sites with potential for formation of G-quadruplex structures is estimated to exceed 300 000 (14,15). Some of these sites are within genes, and potential for G4 DNA formation in genes correlates with gene function; in particular, proto-oncogenes are G-rich and tumor suppressor genes are depleted for G-runs relative to the genomic average (16). Promoter regions are also G-rich (17–19). To understand if and how G-richness and G-quadruplex formation might contribute to regulation of gene expression, it is important to take into account contributions of well-defined mechanisms unrelated to G-quadruplex formation that target individual guanines or runs of guanines. The CpG dinucleotide, site of regulatory methylation, is enriched in 'CpG islands' within promoters (20,21). Some very common transcription factors recognize G-rich sites in duplex DNA, including SP1 (RGGCGKR), KLF (GGGGTGGGG), EKLF (AGGGTGKGG), MAZ (GGGAGGG), EGR-1 (GCGTGGGCG) and AP-2 (CGCCNGSGGG) (22–24). And, during pre-mRNA processing, G-rich sites in RNA are recognized by the hnRNP A (UAGGGU/A) (25) and hnRNP H family (GGGA) proteins (26). It is also important to distinguish between the potential contributions of the template (transcribed) and nontemplate DNA strands. Downstream of the transcription start site (TSS), G-rich regions in the nontemplate strand will become part of the pre-mRNA or (if outside of introns) the mRNA, where they could in principle provide structural targets for regulation. Moreover, transcription of G-rich regions, either *in vitro* or intracellularly, readily produces characteristic structures, G-loops, containing a stable cotranscriptional RNA/DNA hybrid on the template strand and G4 DNA interspersed with single stranded regions on the nontemplate strand (27–29). G4 DNA in the nontemplate strand of a G-loop could similarly provide a regulatory target. In either case, elements would be predicted to display strand bias, and to be concentrated in the nontemplate DNA strand.

Thus far, there has been no systematic effort to integrate our understanding of well-recognized regulatory mechanisms or transcriptional strand bias with the potential for G-quadruplex formation within promoter regions. We have now compared G-richness that predicts potential for DNA or RNA G-quadruplex formation in the upstream and downstream regulatory regions of the 18 217 human RefSeq genes (NCBI 36). We examined the 2 kb flanking the TSS, including 1 kb of upstream sequence (−1000 to −1) and 1 kb of downstream transcribed sequence (+1 to +1000), distinguishing the contributions of the template and nontemplate strands. As documented by others (17–19), we found that regions both upstream and downstream of the TSS are G-rich. However, upstream of the TSS, much G-richness and potential for G4 DNA formation derives from the presence of well-defined canonical regulatory motifs, including CpG dinucleotides, sites of regulatory cytosine methylation; and motifs recognized by the transcription factor SP1. Downstream of the TSS, we identified G-rich elements on the nontemplate DNA strand which were not eliminated by masking G-rich motifs for known factors which bind either DNA or RNA, or by masking CpG dinucleotides. These elements map to the 5′-most 100 bp of the first intron. These elements are in the nontemplate strand, and could therefore be recognized in either the pre-mRNA transcript or in DNA that has been transiently denatured during transcription. Examination of first intron sequences from the genomes of other organisms showed that these G-rich elements are conserved in mouse, chicken, and frog, but not zebrafish. G-rich elements at the 5′-end of intron 1 may provide structural targets for regulation of gene expression at the level of transcription or RNA processing.

METHODS

Sequence data

Sequence data for the 18 217 human RefSeq genes (NCBI 36 assembly) were downloaded from the Ensembl database 46 using BioMart (30,31). From BioMart, we obtained the complete gene sequences, as well as flanking sequences extending 5 kb upstream and downstream of the gene sequences. The regulatory regions analyzed included the 1 kb of sequence flanking TSSs. Intergenic regions included 2 kb of sequence which spanned from 3 to 5 kb upstream of the 5′ ends and downstream of the 3′ ends of the 18 217 RefSeq genes. From Biomart, we also obtained cDNA sequences and coding sequences for the human RefSeq genes.

Intron sequences do not have an Ensembl identifier; therefore, intron sequences were derived from the transcript sequences downloaded from Ensembl database 46 using BioMart (30,31), along with the transcript start and end positions, exon start and end positions, exon rank, and transcribed strand orientation. The first, second and third introns (between exons with rank 1 and 2, 2 and 3, or 3 and 4, respectively) were extracted from the transcript sequences. Each unique intron was distinguished by its Ensembl gene identifier and sequence length, yielding 18 222 first introns, 16 930 second introns and 15 466 third introns for the human genome. Some of these intron sequences may derive from different transcripts of the same gene. First intron sequences were similarly derived from the mouse (Mus musculus, NCBIM 36, 18543 RefSeq genes), chicken (Gallus gallus, WASHUC 2, 4782 known protein coding genes), frog (Xenopus tropicalis, JGI 4.1, 5530 known protein coding genes), and zebrafish (Danio rerio, ZFISH 7, 10578 RefSeq genes) genomes, using transcripts downloaded by BioMart from Ensembl 46.

Control sequences

Pseudo-coding sequences were generated by using the online Sequence Manipulation Site (32). Using the 'Random Coding DNA' applet, we generated 1000 pseudo-coding sequences 3 kb in length, and made up of random codons as defined by the NCBI standard code. To generate random sequences with the same GC content as typical human cDNA, coding or intron sequences, we

shuffled each of the sequences of the human RefSeq gene sequences for each type of sequence. To shuffle the sequences, we used a random number generator to provide an index into each source sequence, deleting each randomly chosen base from the original sequence as it was moved into the newly assembled shuffled sequence.

Sequence analysis

For intramolecular G4 DNA to form within a single-stranded region, four runs of three or more consecutive guanines must be in some proximity, but the limits of this proximity are not known. We elected to evaluate sequences in 100 bp intervals, which represents half the typical spacing between nucleosomes, and is therefore a relatively short region in genomic terms. We thus define a 'G-run' as three or more consecutive guanines, and 'G-richness' as any 100-bp sequence that contains four or more G-runs. We counted G-runs in 100-bp intervals for both nontemplate and template strands of DNA sequences. This analysis separately tallied G-runs within 100-bp intervals, and omitted from the tally G-runs split between two adjacent non-overlapping intervals. This analysis identifies locations in each sequence where the potential for G4 DNA formation exists, thus differing from calculation of potential for G4 DNA formation, 'G4P' (16), which yields one value for an entire sequence.

Regulatory motifs

Conserved regulatory motifs for transcription factors that bind duplex DNA were obtained from the TRANSFAC database 7.0 (24). The following SP1 sequences from consensus RGGCGKR were chosen for analysis because of their frequency and their potential effect on calculations of G-richness: GGGGCGGGG, GGGCGGG, AGGCGGG, GGGCGTG, GGGCGGA. Additional transcription factor binding sites analyzed were: KLF (GGGGTGGGG), EKLF (AGGGTGKGG), MAZ (GGGAGGG), EGR-1 (GCGTGGGCG) and AP-2 (CGCCNGSGGG). G-rich motifs for RNA binding factors included in the analysis were those for the hnRNP A (UAGGGU/A) (25); and hnRNP H family (GGGA) proteins that includes hnRNP H/H'/F/2H9 (26), which we will refer to collectively as hnRNP H motifs.

Masking regulatory motifs

The transcription factor motifs (above) and their reverse complements were masked by converting them to N's, as were CpG dinucleotides, to eliminate them from both DNA strands. The motifs for hnRNP proteins were masked only on the nontemplate strand that corresponds to the mRNA sequence. Some of these motifs may overlap within a sequence, and thus when multiple motifs are evaluated the order of masking motifs can affect the results. For example, the sequence GGGCGGG would be completely masked upon scoring SP1 motifs (GGGCGGG), but if CpG dinucleotides were masked first, the SP1 motif would be eliminated and only the 5′-most GGG motif retained. To minimize this, multiple motifs were masked in the following order: SP1, MAZ, KLF, EKLF, EGR-1, AP-2, hnRNP A, hnRNP H, CpG.

Statistical analysis

The Gaussian curve fits were determined with OriginPro v7.5, and correlation reported by the R^2 value. Spearman correlation was calculated by using the R 2.4.0 statistics package. To determine the significance of strand bias, we compared the numbers of G-runs in a 100-bp interval between the template and nontemplate strands by a two-tailed, paired *t*-test performed with Excel 2003. This analysis establishes if a group of sequences is significantly more G-rich on the template or nontemplate strands. Individual genes can always be exceptions.

RESULTS

The nontemplate strand is G-rich downstream of the TSS

Four 'G-runs' (three or more consecutive guanines) are typically required for the formation of a G-quadruplex, so the density of G-runs provides one measure of potential for G-quadruplex formation. We analyzed the density of G-runs within the promoter regions of all 18 217 RefSeq genes (NCBI 36), examining the 2 kb region surrounding the TSS (range −1000 to +1000). We counted G-runs in 100-bp intervals throughout this region, and calculated the 'G-richness' (four or more G-runs within a span of 100 bp). This confirmed that regions upstream and downstream of the TSS are G-rich (Figure 1A), as documented by others (17–19). We identified two peaks of G-richness, one upstream (−100 to +1) and one downstream (+200 to +300) of the TSS. Within the upstream peak, 55% of genes are G-rich; and within the downstream peak, 57% of genes are G-rich (Figure 1A).

To determine if G-richness exhibits strand bias, the nontemplate and template DNA strands were analyzed separately (Figure 1B). Both strands contribute to both peaks of G-richness. Upstream of the TSS, the two DNA strands are comparably G-rich, and at the peak (−100 to +1) a slightly greater fraction of genes are G-rich on the nontemplate strand (29% and 26%, respectively). Downstream of the TSS, a considerably greater fraction of genes are G-rich on the nontemplate strand. The peak of G-richness on the nontemplate strand is in the region +200 to +300 (Figure 1B). Within this region, 34% of genes are G-rich on the nontemplate strand and 24% on the template strand. Analyses of strand bias by paired *t*-tests for each 100 bp interval showed that the differences at both peaks are significant (Supplementary Figure 1), with significance especially high in the interval +200 to +300 ($P = 1E$–166). It is important to emphasize that the significance between what appear to be small differences (e.g. interval −100 to TSS, $P = 2E$–7) is greatly influenced by the large number (>18K) of sequences included in the analyses. The strand bias downstream of the TSS was especially intriguing, because G-richness of the nontemplate strand within a transcribed region could permit formation of G-quadruplexes in either the newly synthesized RNA transcript or the DNA within a G-loop.

For comparison, we analyzed the profile of G-richness of DNA sequences from two other sources not likely to be enriched in regulatory sequences. One sequence set was

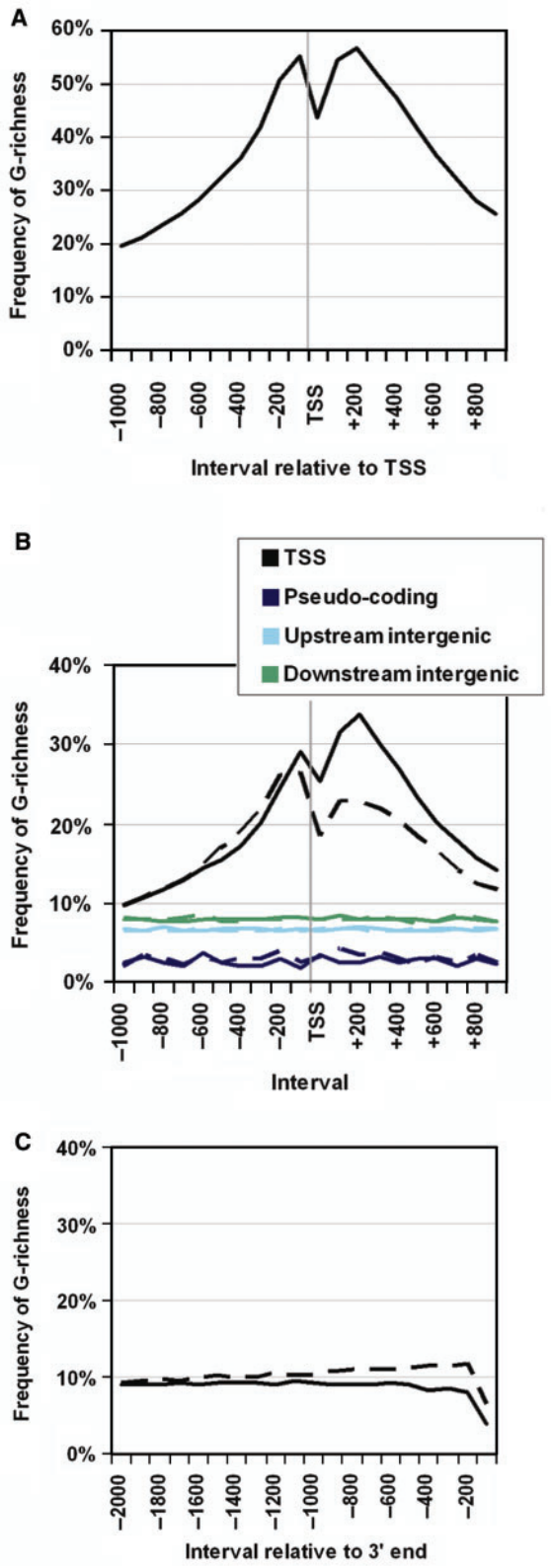

Figure 1. Strand-biased G-richness in human genes. Percentage of genes with four or more G-runs per 100 bp interval was calculated for the indicated regions: (**A**) G-richness of duplex DNA within the 2 kb window spanning the TSS; analysis includes 18 217 human RefSeq genes. (**B**) Strand bias of G-richness. Nontemplate strands (solid lines) and template strands (dashed lines) of human RefSeq genes (black); 1000 random pseudo-coding sequences (blue); intergenic sequences 3 kb upstream of the TSS (cyan); and intergenic sequences 3 kb downstream of the 3′ ends of the genes (green). G-richness of nontemplate and template strands is indistinguishable within intergenic sequences. (**C**) Strand bias of G-richness within 2 kb of the 3′ ends of genes. Nontemplate strands (solid line) and template strands (dashed line).

derived from 1000 pseudo-coding sequences composed of random codons. Only 3% of the pseudo-coding sequences were G-rich, and G-richness was evenly distributed along the 2 kb of DNA analyzed, with no evident strand bias (Figure 1B, blue lines). The other sequence set was derived from 2 kb intergenic regions of the human genome which mapped 3 kb away from either the 5′ or 3′ ends of the 18 217 RefSeq genes. Only 7% and 8% of the upstream and downstream intergenic regions, respectively, were G-rich; and there were neither peaks of G-richness nor any strand bias (Figure 1B, cyan and green lines, respectively).

We also examined the 3′ termini of the human RefSeq genes, separately analyzing the nontemplate and template strands. This analysis showed that 3′ termini are much less G-rich than the region surrounding the TSS (Figure 1C). Furthermore, within the range −2000 to −1500 upstream of the 3′ termini, where 10% of the genes exhibit G-richness on either strand, G-richness does not differ significantly between the two strands ($P > 0.01$). Within the 3′ terminal 1500 bp, fewer genes are G-rich on the nontemplate strand than the template strand; and in the terminal 100 bp, the fraction of genes that are G-rich on the template and nontemplate strands drops to 6% and 4%, respectively. Thus, at the 3′ end of genes, there is not strong potential for formation of G-quadruplex structures or G-loops. Quadruplex structures are therefore in general unlikely to contribute to regulation at the 3′ UTR, either within mRNA transcripts or DNA templates, although specific genes may always be exceptions.

G-richness near the TSS is not due to enrichment of CpG dinucleotides

Many TSSs are embedded in CpG islands (20), and the CpG dinucleotides within these islands are targets of cytosine methylation usually associated with transcription suppression (21,33). To assess the contribution of CpG dinucleotides to G-richness near the TSS, we determined the frequency of genes that are G-rich with all CpG dinucleotides masked (Figure 2A, black lines) and the frequency that can be attributed to CpG dinucleotides (Figure 2A, blue lines). A plot of the contribution of CpG dinucleotides to nontemplate strand G-richness conformed to a nearly perfect Gaussian distribution ($R^2 = 0.99$) centered in the range +100 to +200 (Figure 2A, solid blue line). Therefore masking CpG dinucleotides did not alter two characteristic features of the profile near the TSS: the presence of two peaks of G-richness on either side of the TSS, or the strand-biased G-richness of the nontemplate strand downstream of the TSS (Figure 2A, black lines). Thus, G-richness within regions spanning TSS cannot be explained solely by the density of CpG dinucleotides in this region.

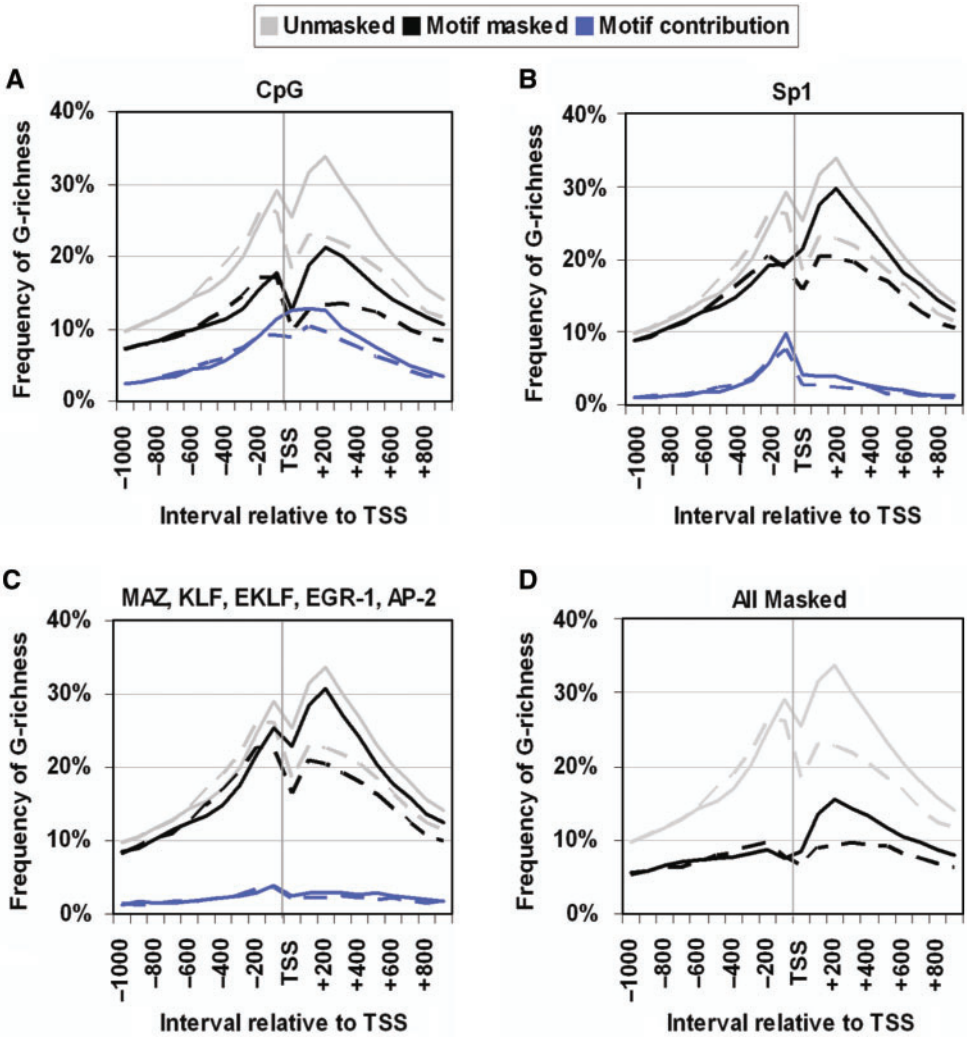

Figure 2. G-richness upstream but not downstream of the TSS can be attributed to canonical regulatory motifs in duplex DNA. Percentage of genes in which G-richness of nontemplate (solid lines) and template (dashed lines) strands was contributed by specific motifs was analyzed for all 18 217 human RefSeq genes within the 2 kb window spanning the TSS. In each panel G-richness of unmasked sequences is shown for comparison (gray). Motifs tested were: (**A**) G-richness contributed solely by CpG dinucleotides (blue); G-richness calculated with CpG dinucleotides masked (black). Gaussian fit (data not shown) for nontemplate strand G-richness contributed by CpG dinucleotides only, represented by the solid blue line ($R^2 = 0.99$). (**B**) G-richness contributed solely by SP1 motifs (blue); G-richness calculated with SP1 motifs masked (black). Gaussian fits (data not shown) for nontemplate strand, SP1 motifs only ($R^2 = 0.80$); and for SP1 motifs masked ($R^2 = 0.95$). (**C**) G-richness contributed by motifs for 5 transcription factors, MAZ, KLF, EKLF, EGR-1, and AP-2 (blue); G-richness with these 5 transcription factor motifs masked (black). (**D**) G-richness with CpG dinucleotides and motifs for transcription factors SP1, MAZ, KLF, EKLF, EGR-1 and AP-2 masked (black).

Motifs for SP1 contribute to G-richness upstream but not downstream of the TSS

To ask how other canonical regulatory motifs might contribute to the overall G-richness of promoters, we next assessed the contributions of motifs for transcription factors that recognize duplex DNA. Duplex DNA-binding sites for transcription factors map both upstream and downstream of the TSS (34). The TRANSFAC database (24), which compiles transcription factors and the sequence motifs that they recognize, lists a number of G-rich consensus motifs for factors that bind duplex DNA. Among the most common regulatory motifs in mammalian genomes is GGGCGGG, recognized by the transcription factor SP1 (RGGCGKR) (22,23). Other common G-rich regulatory motifs recognized by transcription factors include those for KLF (GGGGTGGGG), EKLF (AGGGTGKGG), MAZ (GGGAGGG), EGR-1 (GCGTGGGCG) and AP-2 (CGCCNGSGGG) (22–24).

To establish whether SP1 motifs might contribute to G-richness near the TSS, we searched for SP1 motifs in the 2 kb window flanking the TSS (−1000 to +1000). SP1 motifs contribute significantly to G-richness and account for approximately one third of the G-richness in the region −100 to −1 (Figure 2B). This echoes the position bias identified for SP1 at −63, just upstream of the TSS (23). Moreover, masking SP1 motifs nearly eliminated the peak of genes which are G-rich in the template strand upstream of the TSS (Figure 2B, solid black line). However, masking SP1 motifs did not eliminate the peak of genes

which are G-rich in the nontemplate strand downstream of the TSS (+200 to +300). Within this region, 30% of genes are G-rich even after eliminating the contribution of potential SP1 sites (Figure 2B, solid black line).

To ensure that these results were robust, we duplicated the analysis using the 'Quadparser' software (15). Quadparser employs a different algorithm for identifying motifs with potential for G-quadruplex formation, scoring motifs based on the presence of four or more G-runs separated by 'loops' containing from one to seven nucleotides that may also include G. A motif with more than four G-runs would be counted as one potential quadruplex (provided that it meets the loop criteria), producing a lower estimate of potential quadruplex structures than by our calculation of G-richness. The results from Quadparser reproduced the results of our analyses of G-richness, either with or without SP1 sites masked (Supplementary Figure 2).

Analysis by either our software or by Quadparser showed that the distribution of nontemplate strand G-richness approximates a Gaussian distribution from −1000 to +1000, but with a dip just downstream of the TSS. The evidence that SP1 sites contribute considerably to G-richness upstream of the promoter suggests that the distribution around the TSS might be better represented as the sum of two Gaussian distributions, one contributed by SP1 motifs and centered upstream of the TSS ($R^2 = 0.80$; Figure 2B, solid blue line), and another not associated with SP1 motifs and centered downstream of the TSS ($R^2 = 0.95$; Figure 2B, solid black line).

Motifs for transcription factors other than SP1 make minor contributions to promoter G-richness

Common transcription factors other than SP1 also recognize G-rich motifs in duplex DNA. The most prominent of these factors are KLF, EKLF, MAZ, EGR-1 and AP-2 (22,24). Graphical representation shows that G-richness contributed by motifs for these five factors in the 2 kb window spanning the TSS is low, and contains a very minor peak (−100 to −1; Figure 2C, blue lines). Correspondingly, masking these five motifs only modestly diminished G-richness upstream of the TSS, and had even less effect on G-richness downstream of the TSS (Figure 2C, black lines). Thus, these five motifs appear to make a minor contribution relative to the SP1 motif, as would be predicted by the relative frequencies of motifs for SP1 and these five other factors established by Xie *et al*. (23).

Masking CpG dinucleotides, as well as the six G-rich motifs for transcription factors SP1, KLF, EKLF, MAZ, EGR-1 and AP-2, flattened the distribution of G-richness upstream of the TSS, so that fewer than 10% of genes remained in the class identified as G-rich (Figure 2D). This approaches the level of G-richness identified within nonregulatory regions, but is still above that of random codons (Figure 1B). The effect of masking was comparably evident when applied to specific genes. Elements with potential for quadruplex formation have been identified upstream of the TSS in the MYC, KIT and VEGF proto-oncogenes, prompting speculation that G-quadruplex formation by these elements might provide highly specific targets for regulation, since both structure and sequence could contribute to making them unique within the genome (35–39). However, potential for quadruplex formation at the prototype element in each of these genes was eliminated upon masking canonical regulatory sites (Supplementary Figure 3).

Mapping G-richness to functional regions within genes

The analysis above showed that enrichment of common motifs for transcriptional regulation, particularly CpG dinucleotides and SP1 binding motifs, could explain the peak of G-richness of the nontemplate strand upstream of the TSS but not downstream of the TSS (Figure 2). We therefore sought to identify other functional elements that might account for G-richness downstream of the TSS. To map G-richness within transcribed regions of human genes, we separately examined cDNA sequences, corresponding to all exons in the mature mRNA; coding regions, corresponding to all sequences between the ATG start site and the stop site for translation; and the first and second introns (Figure 3A). Within each of those regions, two subregions were separately analyzed: the 1 kb at the very 5′-end (i.e. just downstream of the TSS of cDNAs, the ATG start codon for coding regions, or the 5′ splice site for introns), and the 1 kb at the very 3′-end (i.e. just upstream of the polyA site of cDNAs, the stop codon for coding regions, or the 3′ splice site for introns). We restricted the analysis of cDNAs, coding regions and introns to regions at least 1 kb in length, which includes more than 11 000 unique sequences for each group; sequences less than 2 kb were included in analysis of both the 5′ and 3′ subregions. Significance of strand bias was determined by paired *t*-tests as in Supplementary Figure 1.

The frequency of genes with G-rich cDNAs is maximal at the very 5′ end, drops precipitously within the first few hundred basepairs, and drops further at the very 3′-end (Figure 3B, green lines). Just downstream of the TSS, 22% of genes are G-rich on the template strand, and 18% are G-rich on the nontemplate strand ($P < 10^{-16}$). A shift in strand bias is evident after the first 100 bp, resulting in a small but significant bias ($P < 10^{-16}$) toward G-richness of the template strand. As the median length of 5′ UTR sequences is greater than 100 bp (156 bp), in most cases G-richness at the very 5′-end of the nontemplate strand maps to 5′-UTRs.

Analysis of coding sequences showed that only 11% can be classified as G-rich at the 5′-end of the nontemplate strand (Figure 3B; blue lines). No significant strand bias is evident in the 5′-most 100 bp region ($P = 0.9$); while further downstream, there is a small but significant bias ($P < 10^{-16}$) toward G-richness of the template strand.

In contrast, analysis of first introns showed that nearly half (48%) are G-rich on the nontemplate strand within the 5′-most 100 bp (Figure 3C, solid red line). There is clear strand bias: G-richness is concentrated on the nontemplate strand, and only half as many first introns are G-rich on the template strand (24%; $P < 10^{-16}$; Figure 3C, red lines). G-richness is concentrated in the 5′-most region of first introns, but the strand bias is

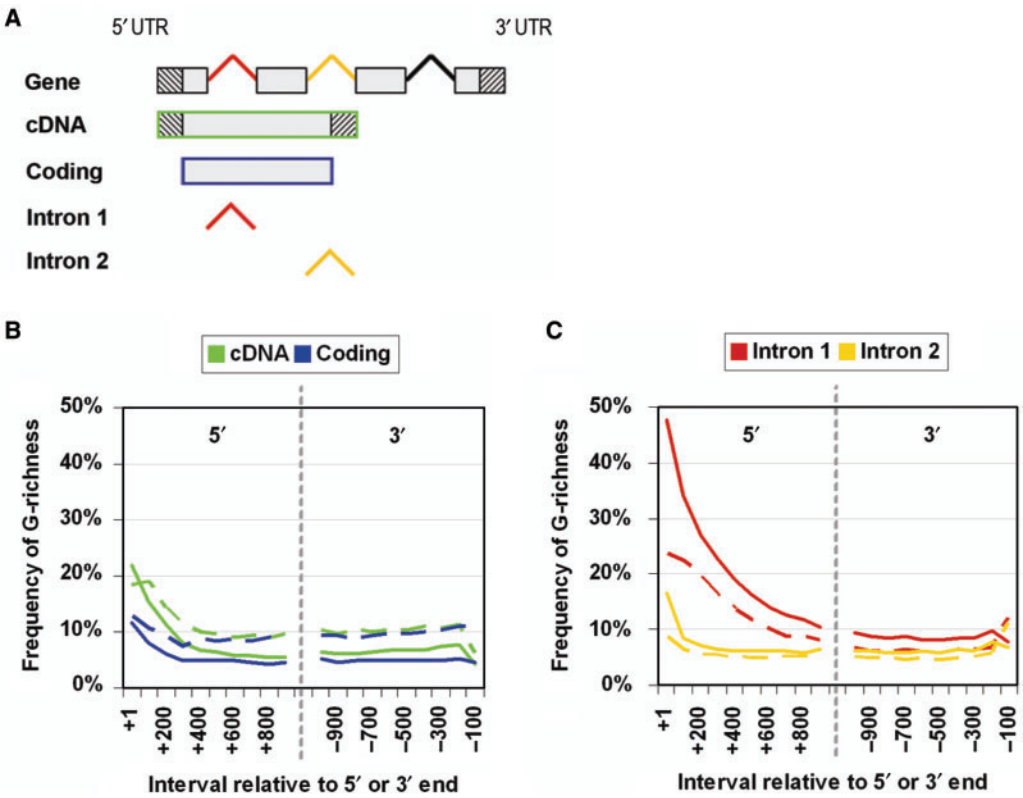

Figure 3. G-richness mapped to functional regions within human genes. (**A**) Diagram of a prototype gene with 5′ UTR (reverse hatched boxes), coding exons (gray boxes), introns (carats), and 3′ UTR (forward hatched boxes) indicated. (**B**) G-richness of 19 056 unique cDNA sequences (green), and 13 640 unique coding sequences (blue). G-richness was calculated for the first 1 kb of each sequence relative to the 5′ end, and the last 1 kb of each sequence relative to the 3′ end, for specific elements of a typical gene, for all sequences greater than 1 kb in length, and distinguishing nontemplate (solid lines) and template (dashed lines) strands. Vertical lines separate analyses of 5′ and 3′ regions. (**C**) G-richness of 13 433 unique first intron sequences (red), and 11 540 unique second intron sequences (gold). Analyses and notations as in (C).

evident throughout, in both 5′- and 3′-intronic regions; and absent at the very 3′-end, near the 3′ splice site. Second introns are characterized by a similar profile of nontemplate strand-biased G-richness (Figure 3C, gold lines), but a lower percentage of second introns than first introns are G-rich at the 5′ end (16% compared to 48%). The profile for third introns is similar to that of second introns (not shown), and the profile for genes that have only one intron is similar to that of all first introns (data not shown). We conclude that G-rich regions at the 5′-end of the first intron, present in 48% of genes, constitute the major component of the strand-biased G-richness evident in RefSeq genes (Figure 1B). In addition, G-rich regions are also present near the 5′-end of cDNA sequences of 22% of genes, where they may in many cases map to the 5′-UTRs. G-rich regions in 5′-UTRs would have the potential to form G-quadruplex structures in mature mRNA transcripts and contribute to regulation of translation, as has been previously suggested (40,41).

Motifs for hnRNP proteins and CpG dinucleotides contribute to but do not account for G-richness of first introns

The bias toward G-rich nontemplate DNA strands in the first intron suggests that G-quadruplex structures may form in this region either as DNA or as RNA. Two hnRNP proteins involved in RNA processing recognize motifs containing runs of three or more guanines in single-stranded DNA or RNA, hnRNP A (UAGGGU/A) and hnRNP H (GGGA) (25,26). These motifs account for more than half of G-rich first introns (Figure 4A, left, compare black and gray lines). However, it is important to note that, in contrast to sites for sequence-specific duplex DNA binding proteins, these motifs may not be sufficient for binding by hnRNP A or hnRNP H, so this calculation almost certainly over-estimates the contribution of motifs for these factors. Despite that, masking of these motifs did not eliminate the peak of G-richness at the very 5′-end of the first introns, as 28% of genes were G-rich even after masking (Figure 4A, left, black line). CpG dinucleotides were the major source of G-richness in the +100 to +200 region downstream of the TSS (Figure 2A, blue line), a region which could overlap with the 5′-end of the first intron. Masking CpG dinucleotides and hnRNP A and hnRNP H motifs eliminated more G-rich genes; but even after this stringent masking, 16% of genes are classified as G-rich at the 5′-end of intron 1 (Figure 4A, left, green line). Moreover, the profile of G-richness was little changed, with a high concentration of G-richness at the very 5′-end.

Motifs for hnRNP A and H similarly account for about half of genes with G-rich second introns (Figure 4A, right,

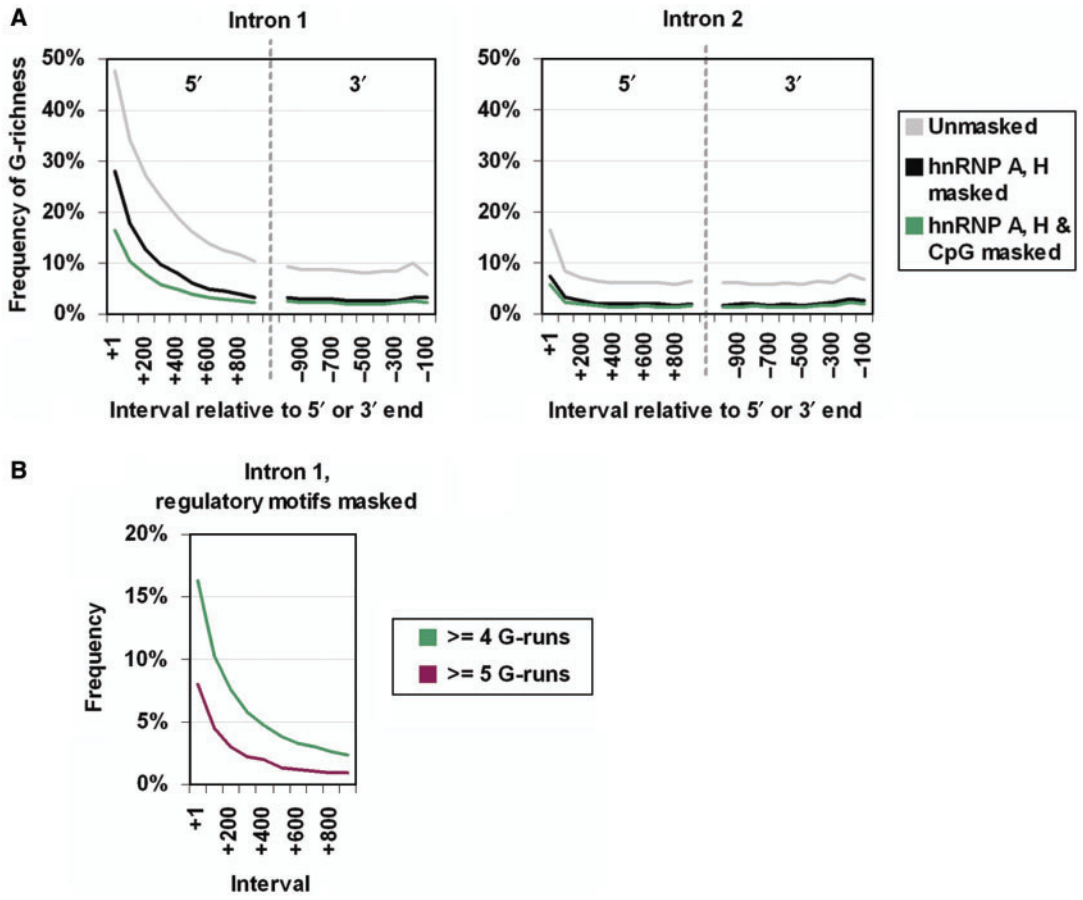

Figure 4. hnRNP A and hnRNP H motifs and CpG dinucleotides contribute to but do not account for G-richness of human first introns. (**A**) Percentage of 13 433 unique first intron sequences (left) or 11 540 unique second intron sequences (right) in which G-richness of nontemplate (solid lines) strands was contributed by specific motifs, within the first 1 kb of each sequence relative to the 5′ end, and the last 1 kb of each sequence relative to the 3′ end, for all sequences that are greater than 1 kb in length. G-richness of unmasked sequences (gray) is shown for comparison with G-richness with motifs for hnRNP A and hnRNP H masked (black), and hnRNP A and hnRNP H plus CpG dinucleotides masked (green). Vertical lines separate the 5′ and 3′ analyses. (**B**) Multiplicity of G-runs in first intron sequences with motifs for hnRNP A and hnRNP H and CpG dinucleotides masked. G-richness with four or more G-runs (green) as in (A), and G-richness redefined as five or more G-runs (plum).

compare black and gray lines). With hnRNP A and hnRNP H motifs masked, masking CpG dinucleotides had very little additional effect on second intron sequences (Figure 4A, right, green line).

Increased multiplicity of G-runs would increase potential for formation of polymorphic G-quadruplexes. We therefore determined the fraction of genes with five or more G-runs per 100 nt in the 5′-most 1 kb of the first intron sequences, with motifs for hnRNP A, hnRNP H, and CpG dinucleotides masked. We found that 8% of genes have five or more G-runs at the 5′-end of intron 1 (Figure 4B, plum line) compared to 16% with four or more G-runs (Figure 4B, green line).

Elements with high potential to form polymorphic G-quadruplex structures at the 5′-end of first introns

Results above demonstrate that first introns of 48% of human genes have potential to form G-quadruplex structures, as they contain four or more G-runs per 100 nt on the nontemplate strand. In addition, half of the G-rich first introns have the potential to form polymorphic G-quadruplexes, even after masking for known regulatory motifs. As increased multiplicity of G-runs would increase combinatorial potential for formation of G-quadruplexes, we determined the frequency of numbers of G-runs per 100 nt of the nontemplate strand in first and second introns, as well as in cDNA and coding regions (Figure 3A) as controls. This analysis evaluated all such sequences >100 bp in length. To provide comparison of observed and predicted values for sequences of identical base composition, the distribution of the numbers of G-runs was calculated for the same groups of sequences randomly shuffled. This analysis showed that all sequences are depleted for low numbers of G-runs (zero or one) in the region +1 to +100 as compared to their corresponding shuffled sequences, with the difference between observed (Figure 5A, bars) and predicted (Figure 5A, lines) values most striking in first introns (red). There is a modest enrichment of the frequency of cDNAs and second introns with four or more G-runs, while coding sequences conform very closely to the expected distribution. Strikingly, the frequency of first introns containing four or more G-runs is considerably greater than predicted

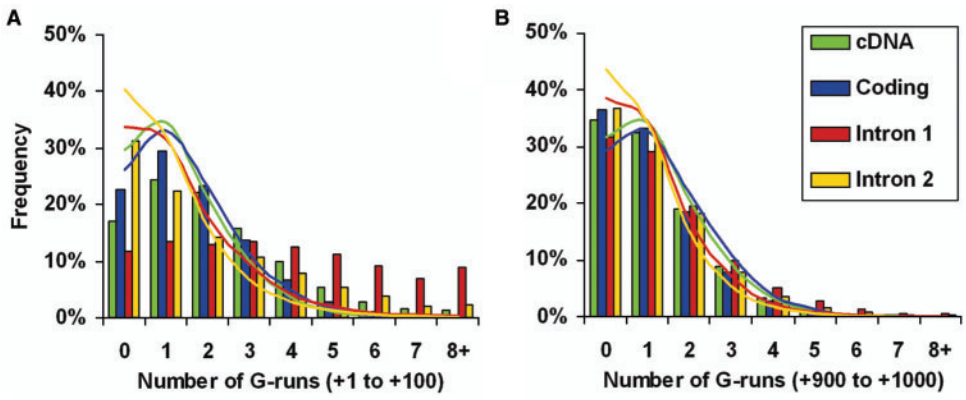

Figure 5. The G-rich element at the 5′ end of first introns has high potential to form polymorphic G-quadruplex structures. Numbers of G-runs were enumerated in 100 nt intervals within the nontemplate strand for each specific element of a typical gene (Figure 3A), including cDNA (green), coding (blue), first intron (red) and second intron (gold), for all sequences greater than 100 bp in length. The distribution of numbers of G-runs is shown for two intervals, comparing the observed value of each genomic region (bars) to the value predicted based upon analysis of the same sequences randomly shuffled (lines). Intervals analyzed were: (**A**) 100 nt interval from +1 to +100 relative to the 5′ end. (**B**) 100 nt interval from +900 to +1000 relative to the 5′ end.

based on sequence composition (Figure 5A, compare red bars and line).

In the human genome, GC content of first introns is inversely correlated with intron length (42). To ensure that the results above were applicable to longer introns, we also determined distribution of G-runs in human first introns greater than 1 kb in length. This produced essentially identical results (data not shown) to analysis of first introns >100 bp (Figure 5A).

In addition, we asked if intron length correlates with the number of G-runs in the first 100 bp, but found that it does not (Spearman correlation, $\rho = -0.014$, $P = 0.07$). Thus, in contrast to GC content (42), G-richness and multiplicity of G-runs at the 5′-end of first introns are not a function of intron length.

To compare G-richness at 5′-ends and elsewhere, we carried out an identical analysis of the same panel of cDNA, coding, and first and second intron sequences, but focused on the region from +900 to +1000. Overall, the differences between observed and predicted values were not striking in this region, although a modest enrichment of genes containing four or more G-runs in first intron sequences was evident (Figure 5B, red).

The results above identify elements with high potential for G4 DNA formation in the nontemplate DNA strand at the 5′-end of first introns in the human genome. Notably, the frequencies of genes containing four, five, six, seven or eight or more G-runs in the first 100 nt of intron 1 are nearly comparable: 13%, 11%, 9%, 7% and 9%, respectively. Together, they account for a total of over 8000 human genes, or nearly half the genes in the most current database.

Conservation of G-rich first intron elements

To establish whether the presence of G-rich elements in the first intron is conserved, we evaluated sequences from the mouse, chicken, frog and zebrafish genomes. As in the analysis of human first introns (Figure 3B), we examined the 1 kb comprising the 5′-most region, for all sequences at least 1 kb in length. Significance of strand bias was determined by paired *t*-tests as in Supplementary Figure 1. We found nontemplate strand-biased G-richness at the 5′-end of first introns in mouse, chicken and frog (Figure 6A). The mouse and chicken profiles closely resemble the human profile. The fraction of genes with four or more G-runs per 100 nt at the 5′-end of the first intron is 41% in mouse and 42% in chicken (Figure 6A, left; cf. 48% in human); and clear strand bias is evident, as more than twice as many genes are G-rich on the nontemplate as on the template strand (16% and 12%, respectively; $P < 10^{-16}$; Figure 6A, right). In the frog, G-rich first introns characterize a smaller fraction of genes (Figure 6A); but there is nonetheless clear strand bias, as 14% of genes are G-rich on the nontemplate strand (Figure 6A, left), and only 4% on the template strand ($P < 10^{-16}$; Figure 6A, right). In zebrafish, first introns are not G-rich.

To establish whether the first introns of these genomes have high densities of G-runs, we evaluated the multiplicity of G-runs in the first 100 nt of the nontemplate strand of first introns for all sequences greater than 100 bp in length (Figure 6B). We found that, in mouse and chicken, 28% and 33% of first introns contain five or more G-runs in the first 100 nt, respectively, very comparable to human (36%). Thus, the vast majority of these regions (human, 75%; mouse, 68%; chicken, 79%) have the potential not only for G-quadruplex formation but for formation of polymorphic G-quadruplex structures. In frog, 7%, or half of the G-rich first introns, have five or more G-runs in the first 100 nt. Less than 1% of the zebrafish intron sequences are G-rich in the first 100 nt. Therefore, G-rich intron 1 elements appear to have evolved around the time of divergence of fish and frogs, and have been widely adopted in human genes. Moreover, potential for formation of not only G-quadruplexes, but polymorphic quadruplex structures, is conserved.

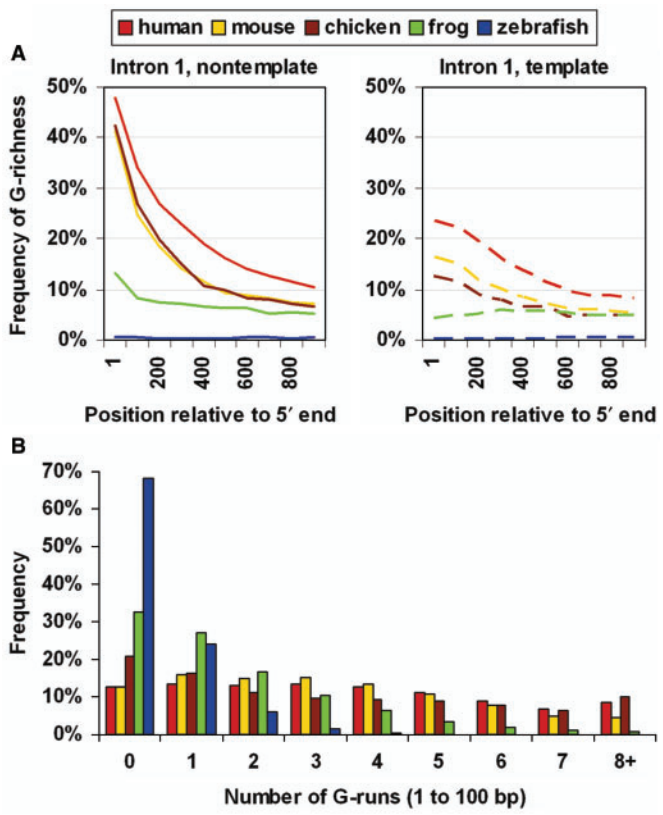

Figure 6. The G-rich element at the 5′ end of first introns is conserved. Comparison of G-richness of the first intron sequences of human (red), mouse (gold), chicken (brown), frog (green) and zebrafish (blue). (**A**) G-richness was calculated for first intron sequences of mouse (11 816), chicken (3399), frog (4193), zebrafish (5787), and compared with human (13 433). Regions analyzed were the 100 nt interval from +1 to +100 relative to the 5′ end, for all unique first introns greater than 1 kb in length, for the nontemplate strand (left, solid lines), and template strand (right, dashed lines). (**B**) Distribution of numbers of G-runs in the first 100 nt of the nontemplate strand of the first intron, for all unique intron sequences greater than 100 bp.

DISCUSSION

We determined the potential for formation of G-quadruplex structures near the TSS of the 18 217 human RefSeq genes, distinguishing the contributions of the template and nontemplate strands in the −1000 to +1000 region surrounding the TSSs. We found that two distinct components determine G-richness. Upstream of the TSS, G-richness is accounted for by well-defined motifs in duplex DNA including CpG dinucleotides for methylation and G-rich motifs recognized by transcription factors that recognize duplex DNA. Downstream of the TSS, G-richness could not be accounted for by CpG dinucleotides or known recognition motifs for factors that bind duplex DNA or RNA. G-rich elements with potential for formation of G-quadruplex structures were found to characterize the nontemplate strand of the 5′-end of first introns of many human genes. We will refer to these G-rich intron 1 elements as 'GrIn1 elements'. Strand bias, position, multiplicity of G-runs and conservation from frogs through humans all suggest that GrIn1 elements may provide structural targets for regulation of gene expression.

Distinct peaks of G-richness upstream and downstream of the TSS

For intramolecular G4 DNA to form, four G-runs must be in proximity, but the limits of this proximity have not been established *in vitro* or *in vivo*. We therefore evaluated sequences 100 bp in length, half the typical spacing between nucleosomes, and scored G-richness as the presence of four or more runs of three or more consecutive guanines within 100 nt on each DNA strand. We established that the results of this analysis are robust with respect to details of the algorithm used for analysis by comparing G-richness as determined by our software and by Quadparser (15). Analyses of G-richness either with our software or with Quadparser produced comparable profiles. In particular, graphic display of G-richness calculated by both methods identified two overlapping peaks, one upstream and one downstream of the TSS. These two peaks of G-richness flanking the TSS appear to represent the superposition of two Gaussian distributions. This possibility was supported by the identification of distinct features that contribute to G-richness upstream and downstream of the TSS.

G-rich regions upstream of the TSS

Our results show that canonical regulatory motifs, including CpG dinucleotides and G-rich duplex motifs for common transcription factors, account for most G-richness and potential for G-quadruplex formation upstream of the transcription start site. The simplest interpretation of our results is that transcriptional regulation at sites upstream of the TSS is determined by canonical regulatory mechanisms acting on duplex DNA. However, we recognize that the masking test that we have used may eliminate sites with potential for G-quadruplex formation which resemble motifs for duplex DNA-binding transcription factors, but do not function as such. Thus, our analysis does not exclude the possibility that G-quadruplexes could contribute to regulation at specific promoters; or that a transformation of duplex to G-quadruplex conformation at a specific site could prevent factor binding and thus alter regulation, as has been proposed to occur at an SP1 site at the VEGF promoter (38). Nonetheless, our analysis does provide a message of caution about drawing functional implications from the presence of G-rich regions within promoters. For example, results of mutagenesis of promoter sequences must be interpreted cautiously, as mutations intended to interfere with potential G-quadruplex formation (e.g. 35) could impair factor binding. To take another example, hypersensitivity to DNase1, which cleaves in the minor groove of duplex DNA, is well known to reflect relaxation of chromatin coincident with binding of factors that activate transcription (34). A correlation has been noted between DNase1 hypersensitivity and sites with potential for G-quadruplex formation, and interpreted as evidence of G4 DNA formation (17).

Our analysis provides the alternative explanation that G-rich regions are quite generally sites of transcription factor binding.

G-richness and potential for G4 DNA formation at specific human promoters has given rise to speculation that G-quadruplex structures might serve as targets for regulation (43–46). These speculations were supported by analysis of structure formation in either synthetic oligonucleotides or supercoiled plasmid DNAs bearing the sites of interest. While G-quadruplexes do form readily in these substrates, and exhibit considerable thermodynamic stability once formed, analyses have yet to be undertaken to establish how normally duplex genomic DNA upstream of the promoter would be denatured to allow conversion into a quadruplex structure; to determine the role of nucleosomes in promoting or inhibiting G-quadruplex formation; or to learn whether these structures withstand attack by G4 DNA helicases or other factors (6–10) which maintain genomic structure *in vivo*. We note in particular that masking CpG dinucleotides and G-rich motifs for common transcription factors (SP1, KLF, EKLF, MAZ, EGR-1 and AP-2) eliminated potential quadruplexes at sites in three proto-oncogene promoters that have served as paradigms for those approaches: MYC, KIT, and VEGF (36,38,47).

More generally, our results challenge the rationale behind attempts to develop small molecule therapeutics directed at presumptive unique quadruplex structures upstream of the TSS at single copy genes. We suggest that the overall G-richness at promoters documented here and by others (17,19) would create such great potential for combinatorial diversity in structure formation that, even if DNA upstream of the TSS were to become denatured and form G-quadruplexes, it would be very difficult to predict which G-runs participated in G-quadruplex formation; and the identity of quadruplexes that did form could differ from cell to cell. The same considerations would suggest that repetitive G-rich sequences, such as those at the telomeres (48), might be a viable therapeutic target since structural polymorphism would be limited and sequence reiteration would tend to produce many copies of identical structures.

G-rich intron 1 elements (GrIn1) and G-richness downstream of the TSS

The most notable features of G-richness downstream of the TSS are the strand bias; the clustering of G-runs at the 5′-end of intron 1; the multiplicity of G-runs; and the conservation of all these features from frogs to humans. G-richness downstream of the TSS could not be accounted for by CpG dinucleotides or binding sites for factors thus far identified which recognize duplex DNA or RNA. It is of course possible that motifs not yet identified with specific factors account for some or all of the G-richness downstream of the TSS. We note that 'G-triplet' motifs (GGG) have been identified at both the 5′ and 3′-ends of introns and associated with regulation of splicing (49–52), and such motifs could contribute to the G-richness we have identified at the 5′-ends of the intron sequences. The mechanism by which G-triplets enhance splicing is not well-understood. Nonetheless, it appears not to involve G-quadruplex formation in the first intron, as addition of fewer than four G-triplets to an internal intron can affect splicing (51).

G-richness downstream of the TSS characterizes the nontemplate but not the template strand. This strand bias would enable formation of G-quadruplexes in G-loops in the nontemplate DNA strand of a transcribed G-rich region (27–29), or in pre-mRNA. Analysis of nontemplate strand intron sequences alone mapped the peak of G-rich regions within the 100 nt just downstream of the 5′ splice site. This position is consistent with our mapping of the peak in G-richness on the nontemplate strand of all RefSeq genes to the region spanning +200 to +300 (Figure 1B): the median length of first exons is 198 bp, so the observed peak at +200 to +300 is just downstream of first exons, and at the 5′-end of first introns.

The first intron represents a potentially privileged position within genomic sequence, as it is the region that maps closest to the promoter but does not appear in the mature mRNA. Proximity to the promoter would enable the common G-rich elements to determine loading of factors critical for regulation of gene expression by either transcription or splicing. As the intron is eliminated upon splicing, these elements would not be under the sequence and structural constraints that would apply to elements that are part of a mature transcript.

Even with motifs for hnRNP proteins and CpG dinucleotides masked, almost 1500 human genes (8%) contain GrIn1 elements with five or more G-runs. This multiplicity of G-runs beyond the four necessary for quadruplex formation would confer potential for combinatorial polymorphism in G-quadruplex structures. Thus, if G-quadruplexes do provide targets for regulation of gene expression, recognition may depend on structural features rather than sequence motifs. Proteins can recognize G-quadruplexes with nanomolar affinity (8,10,11), comparable to (or better than) that of factors that recognize specific sequence motifs in duplex DNA. Therefore it is intriguing to hypothesize that G-quadruplexes formed by GrIn1 elements provide structural targets for regulation of gene expression. We emphasize that a mechanism that takes advantage of these elements has yet to be discovered, presenting the challenge for future experiments.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

Funding was provided by US National Institutes of Health R01 GM65988 and National Cancer Institute P01 CA77852 to NM; National Cancer Institute CA009537 (Basic and Cancer Immunology Training Grant) and Cancer Research Institute (Tumor Immunology Pre-doctoral Training Grant) to J.E. Funding to pay the

Open Access publication charges for this article was provided by US NIH R01 GM65988.

Conflict of interest statement. None declared.

REFERENCES

1. Phan,A.T., Kuryavyi,V. and Patel,D.J. (2006) DNA architecture: from G to Z. *Curr. Opin. Struct. Biol.*, **16**, 288–298.
2. Maizels,N. (2006) Dynamic roles for G4 DNA in the biology of eukaryotic cells. *Nat. Struc. Mol. Biol.*, **13**, 1055–1059.
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. Rachwal,P.A., Brown,T. and Fox,K.R. (2007) Effect of G-tract length on the topology and stability of intramolecular DNA quadruplexes. *Biochemistry*, **46**, 3036–3044.
5. Rachwal,P.A., Findlow,I.S., Werner,J.M., Brown,T. and Fox,K.R. (2007) Intramolecular DNA quadruplexes with different arrangements of short and long loops. *Nucleic Acids Res.*, **35**, 4214–4222.
6. Sun,H., Karow,J.K., Hickson,I.D. and Maizels,N. (1998) The Bloom's syndrome helicase unwinds G4 DNA. *J. Biol. Chem.*, **273**, 27587–27592.
7. Huber,M.D., Duquette,M.L., Shiels,J.C. and Maizels,N. (2006) A conserved G4 DNA binding domain in RecQ family helicases. *J. Mol. Biol.*, **358**, 1071–1080.
8. 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.
9. Wu,X. and Maizels,N. (2001) Substrate-specific inhibition of RecQ helicase. *Nucleic Acids Res.*, **29**, 1765–1771.
10. 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.
11. Hanakahi,L.A., Sun,H. and Maizels,N. (1999) High affinity interactions of nucleolin with G-G-paired rDNA. *J. Biol. Chem.*, **274**, 15908–15912.
12. Dempsey,L.A., Sun,H., Hanakahi,L.A. and Maizels,N. (1999) G4 DNA binding by LR1 and its subunits, nucleolin and hnRNP D, A role for G-G pairing in immunoglobulin switch recombination. *J. Biol. Chem.*, **274**, 1066–1071.
13. Khateb,S., Weisman-Shomer,P., Hershco,I., Loeb,L.A. and Fry,M. (2004) Destabilization of tetraplex structures of the fragile X repeat sequence (CGG)n is mediated by homolog-conserved domains in three members of the hnRNP family. *Nucleic Acids Res.*, **32**, 4145–4154.
14. Todd,A.K., Johnston,M. and Neidle,S. (2005) Highly prevalent putative quadruplex sequence motifs in human DNA. *Nucleic Acids Res.*, **33**, 2901–2907.
15. Huppert,J.L. and Balasubramanian,S. (2005) Prevalence of quadruplexes in the human genome. *Nucleic Acids Res.*, **33**, 2908–2916.
16. 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.
17. Huppert,J.L. and Balasubramanian,S. (2007) G-quadruplexes in promoters throughout the human genome. *Nucleic Acids Res.*, **35**, 406–413.
18. 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.
19. Zhao,Y., Du,Z. and Li,N. (2007) Extensive selection for the enrichment of G4 DNA motifs in transcriptional regulatory regions of warm blooded animals. *FEBS Lett.*, **581**, 1951–1956.
20. Saxonov,S., Berg,P. and Brutlag,D.L. (2006) A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters. *Proc. Natl Acad. Sci. USA*, **103**, 1412–1417.
21. Strathdee,G., Sim,A. and Brown,R. (2004) Control of gene expression by CpG island methylation in normal cells. *Biochem. Soc. Trans.*, **32**, 913–915.
22. Bina,M., Wyss,P., Ren,W., Szpankowski,W., Thomas,E., Randhawa,R., Reddy,S., John,P.M., Pares-Matos,E.I. *et al.* (2004) Exploring the characteristics of sequence elements in proximal promoters of human genes. *Genomics*, **84**, 929–940.
23. Xie,X., Lu,J., Kulbokas,E.J., Golub,T.R., Mootha,V., Lindblad-Toh,K., Lander,E.S. and Kellis,M. (2005) Systematic discovery of regulatory motifs in human promoters and 3′ UTRs by comparison of several mammals. *Nature*, **434**, 338–345.
24. Matys,V., Kel-Margoulis,O.V., Fricke,E., Liebich,I., Land,S., Barre-Dirrie,A., Reuter,I., Chekmenev,D., Krull,M. *et al.* (2006) TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. *Nucleic Acids Res.*, **34**, D108–D110.
25. Burd,C.G. and Dreyfuss,G. (1994) RNA binding specificity of hnRNP A1: significance of hnRNP A1 high-affinity binding sites in pre-mRNA splicing. *Embo J*, **13**, 1197–1204.
26. Caputi,M. and Zahler,A.M. (2001) Determination of the RNA binding specificity of the heterogeneous nuclear ribonucleoprotein (hnRNP) H/H'/F/2H9 family. *J. Biol. Chem.*, **276**, 43850–43859.
27. 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.
28. 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.
29. Duquette,M.L., Huber,M.D. and Maizels,N. (2007) G-rich proto-oncogenes are targeted for genomic instability in B-cell lymphomas. *Cancer Res.*, **67**, 2586–2594.
30. Hubbard,T., Andrews,D., Caccamo,M., Cameron,G., Chen,Y., Clamp,M., Clarke,L., Coates,G., Cox,T. *et al.* (2005) Ensembl 2005. *Nucleic Acids Res.*, **33**, D447–D453.
31. Durinck,S., Moreau,Y., Kasprzyk,A., Davis,S., De Moor,B., Brazma,A. and Huber,W. (2005) BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. *Bioinformatics*, **21**, 3439–3440.
32. Stothard,P. (2000) The sequence manipulation suite: JavaScript programs for analyzing and formatting protein and DNA sequences. *Biotechniques*, **28**, 1102–1104.
33. Appanah,R., Dickerson,D.R., Goyal,P., Groudine,M. and Lorincz,M.C. (2007) An unmethylated 3′ promoter-proximal region is required for efficient transcription initiation. *PLoS Genet.*, **3**, e27.
34. Birney,E., Stamatoyannopoulos,J.A., Dutta,A., Guigo,R., Gingeras,T.R., Margulies,E.H., Weng,Z., Snyder,M., Dermitzakis,E.T. *et al.* (2007) Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature*, **447**, 799–816.
35. 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.
36. 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.
37. Phan,A.T., Kuryavyi,V., Burge,S., Neidle,S. and Patel,D.J. (2007) Structure of an unprecedented G-quadruplex scaffold in the human c-kit promoter. *J. Am. Chem. Soc.*, **129**, 4386–4392.
38. 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.
39. Hurley,L.H., Von Hoff,D.D., Siddiqui-Jain,A. and Yang,D. (2006) Drug targeting of the c-MYC promoter to repress gene expression via a G-quadruplex silencer element. *Semin. Oncol.*, **33**, 498–512.
40. Wieland,M. and Hartig,J.S. (2007) RNA quadruplex-based modulation of gene expression. *Chem. Biol.*, **14**, 757–763.
41. Kumari,S., Bugaut,A., Huppert,J.L. and Balasubramanian,S. (2007) An RNA G-quadruplex in the 5′ UTR of the NRAS proto-oncogene modulates translation. *Nat. Chem. Biol.*, **3**, 218–221.
42. Gazave,E., Marques-Bonet,T., Fernando,O., Charlesworth,B. and Navarro,A. (2007) Patterns and rates of intron divergence between humans and chimpanzees. *Genome Biol.*, **8**, R21.

43. Hurley,L.H. (2001) Secondary DNA structures as molecular targets for cancer therapeutics. *Biochem. Soc. Trans.*, **29**, 692–696.
44. Patel,D.J., Phan,A.T. and Kuryavyi,V. (2007) Human telomere, oncogenic promoter and 5′-UTR G-quadruplexes: diverse higher order DNA and RNA targets for cancer therapeutics. *Nucleic Acids Res.*, doi:10.1093/nar/gkm711.
45. Todd,A.K., Haider,S.M., Parkinson,G.N. and Neidle,S. (2007) Sequence occurrence and structural uniqueness of a G-quadruplex in the human c-kit promoter. *Nucleic Acids Res.*, **35**, 5799–5808.
46. Huppert,J.L. (2007) Four-stranded DNA: cancer, gene regulation and drug development. *Philos. Trans. A. Math. Phys. Eng. Sci.*, **365**, 2969–2984.
47. Simonsson,T., Pecinka,P. and Kubista,M. (1998) DNA tetraplex formation in the control region of c-myc. *Nucleic Acids Res.*, **26**, 1167–1172.
48. De Cian,A., Lacroix,L., Douarre,C., Temime-Smaali,N., Trentesaux,C., Riou,J.F. and Mergny,J.L. (2008) Targeting telomeres and telomerase. *Biochimie*, **90**, 131–155.
49. McCullough,A.J. and Berget,S.M. (1997) G triplets located throughout a class of small vertebrate introns enforce intron borders and regulate splice site selection. *Mol. Cell. Biol.*, **17**, 4562–4571.
50. McCullough,A.J. and Berget,S.M. (2000) An intronic splicing enhancer binds U1 snRNPs to enhance splicing and select 5′ splice sites. *Mol. Cell. Biol.*, **20**, 9225–9235.
51. Yeo,G., Hoon,S., Venkatesh,B. and Burge,C.B. (2004) Variation in sequence and organization of splicing regulatory elements in vertebrate genes. *Proc. Natl Acad. Sci. USA*, **101**, 15700–15705.
52. Majewski,J. and Ott,J. (2002) Distribution and characterization of regulatory elements in the human genome. *Genome Res.*, **12**, 1827–1836.