Published online 17 July 2009

Nucleic Acids Research, 2009, Vol. 37, No. 17 ***5749–5756***
doi:10.1093/nar/gkp590

The disruptive positions in human G-quadruplex motifs are less polymorphic and more conserved than their neutral counterparts

Sigve Nakken[1,*], Torbjørn Rognes[1,2] and Eivind Hovig[2,3,4]

[1]Centre for Molecular Biology and Neuroscience, Institute of Medical Microbiology, Oslo University Hospital, Rikshospitalet, NO-0027, Oslo, [2]Department of Informatics, University of Oslo, PO Box 1080 Blindern, NO-0316, Oslo, [3]Department of Tumor Biology, Institute for Cancer Research and [4]Department of Medical Informatics, Oslo University Hospital, Norwegian Radium Hospital, Montebello, NO-0310, Oslo, Norway

Received May 18, 2009; Revised June 25, 2009; Accepted June 26, 2009

ABSTRACT

Specific guanine-rich sequence motifs in the human genome have considerable potential to form four-stranded structures known as G-quadruplexes or G4 DNA. The enrichment of these motifs in key chromosomal regions has suggested a functional role for the G-quadruplex structure in genomic regulation. In this work, we have examined the spectrum of nucleotide substitutions in G4 motifs, and related this spectrum to G4 prevalence. Data collected from the large repository of human SNPs indicates that the core feature of G-quadruplex motifs, 5′-GGG-3′, exhibits specific mutational patterns that preserve the potential for G4 formation. In particular, we find a genome-wide pattern in which sites that disrupt the guanine triplets are more conserved and less polymorphic than their neutral counterparts. This also holds when considering non-CpG sites only. However, the low level of polymorphisms in guanine tracts is not only confined to G4 motifs. A complete mapping of DNA three-mers at guanine polymorphisms indicated that short guanine tracts are the most under-represented sequence context at polymorphic sites. Furthermore, we provide evidence for a strand bias upstream of human genes. Here, a significantly lower rate of G4-disruptive SNPs on the non-template strand supports a higher relative influence of G4 formation on this strand during transcription.

INTRODUCTION

Human genomic DNA usually exists in the double-stranded conformation, but during denaturation, single strands containing tandemly repeated sequences can assemble into higher order DNA structures. In repetitive and guanine-rich sequences of the genome, single-stranded DNA can adopt four-stranded structures known as G-quadruplexes or G4 DNA (1). The G-quadruplex comprises a stack of G-tetrads, which are planar arrays of four guanines connected by Hoogsteen hydrogen bonds (2). G-quadruplexes are rapidly stabilized in the presence of monovalent cations, and their folding topology is influenced by the length and composition of short-sequence loops that link the stacked G-tetrads together (3–6). The first *in vitro* observations of G-quadruplex formation came from the single-stranded overhang at human telomeres (7,8), a sequence characterized by tandem repeats of TT AGGG. This finding was later followed by studies that demonstrated the existence of G-quadruplexes *in vivo* (9–11). The hypothesized role of G-quadruplex formation in living cells has received further support from the recognition of conserved factors that selectively bind and unwind G4 (12–15). However, the relative impact of G-quadruplex formation in the context of gene regulation and genome stability is still unclear.

Computational algorithms have been used to scan the human genome for the G4 consensus motif, which is a sequence containing at least four runs of at least three guanines (G-tracts) (16–18). These scans have identified enrichment in a number of chromosomal regions of biological importance, including the ribosomal DNA (19), the immunoglobulin heavy chain switch regions (20), telomeres (21) and transcriptional regulatory regions (22,23). With respect to gene transcription, different modes of G4-mediated regulation have been proposed. In one scenario, the formation of G4 is thought to increase the rate of transcription by preventing renaturation of double-stranded DNA (23). Others have though shown experimentally how small compounds can stabilize a promoter G-quadruplex and thereby decrease the expression rate (24). The idea that G-quadruplexes may act as regulators

*To whom correspondence should be addressed. Tel: +47 22 84 47 86; Fax: +47 22 84 47 82; Email: sigve.nakken@medisin.uio.no

of gene expression has been strengthened by multiple observations of G-quadruplex formation in human promoters, including the proto-oncogenes c-MYC (24,25) and c-KIT (26), as well as muscle-specific genes (27). Moreover, G4 motifs appear to be enriched in the promoters of other warm-blooded animals (28). Within motifs, there is a considerable preference for single-nucleotide loops between the consecutive guanine runs, and this is also characteristic of the experimentally derived structures that are most stable (22,29,30). The latter studies showed how a correlation between common sequence features of G4 motifs and observations *in vitro* might aid the interpretation of G4 prevalence. An important set of data that remains to be explored in this respect is the spectrum of common nucleotide polymorphisms in G4 motifs, and how this spectrum relates to findings from recent kinetic and spectroscopic studies of mutated G4 (31,32). The studies of single-base mutated G-quadruplexes have demonstrated a strong relationship between quadruplex stability and the mutation position, with the central guanines of G-tracts being most critical for stable quadruplex folds. Thus, if the G-quadruplexes exhibit biological activity in genomic regions, one would expect to see a relatively lower rate of polymorphic bases at critical sites of the G4 motif, as a consequence of negative selection. Taking into account the non-randomness of point mutagenesis, in which both base composition and DNA sequence contexts influence substitution rates (33–36), it is therefore of importance to see how the different sites in G4 motifs relate to known genetic variation in the form of human single nucleotide polymorphisms (SNPs). The collection of DNA polymorphisms in G4 motifs also represents an additional dimension in the identification of genomic regions undergoing G4 selection. In particular, the relative rate of G4-disruptive SNPs could indicate the extent of selection for the G-quadruplex structure in different genomic regions.

Here, we report a genome-wide analysis of SNPs in human G-quadruplex motifs, with an emphasis towards their occurrences in gene and regulatory sequences. We have used a large collection of validated SNPs from dbSNP as our data source of nucleotide substitutions (37). Overall, the results demonstrate a non-random pattern of nucleotide polymorphism in G-quadruplex motifs. In particular, we show that the internal sites of guanine runs are well protected from polymorphisms in the human genome, indicating a relationship between sequence-dependent mutagenesis of guanine and the prevalence of guanine tracts.

MATERIALS AND METHODS

SNP data

dbSNP (build 129, released on 18 April 2008) was downloaded in XML format from ftp://ftp.ncbi.nlm.nih.gov/snp/. We included SNPs that (i) were biallelic, (ii) had been uniquely mapped to the human genome with an alignment accuracy of at least 99%, (iii) had been validated by at least one of NCBI's validation criteria (that is, 'by-frequency', 'byCluster', 'by2Hit2Allele' or 'byOtherPop') and (iv) if genotyped by the HapMap project, had a minor allele frequency of at least 1% in minimum one of the sampled populations. A total of 5 717 575 SNPs satisfied the criteria above.

Sequence and annotation data

We used the *quadparser* algorithm to retrieve all sequences in the human genome (NCBI build 36.3) capable of forming a G-quadruplex, identified by the sequence motif $G_{3+}N_{1-7}G_{3+}N_{1-7}G_{3+}N_{1-7}G_{3+}$, where G is guanine and N is any nucleotide (16). This simple consensus was inferred after several biophysical experiments had investigated the sequence basis for stable quadruplex folds (3,4), and represents the most common approach to map the grand total of potential G-quadruplex forming sequences. From the *quadparser* output, we extracted each putative G-quadruplex motif, regardless of any potential overlap with a neighboring motif [this corresponds to the 'un-restricted' set of G4 motifs, as defined by Todd *et al.* (17)]. Motifs with guanine tracts of length greater than six were excluded. The choice of overlapping motifs allowed us to evaluate the context and effect of a SNP for each individual putative G-quadruplex-forming structure. We only considered SNPs that mapped to G4 motifs present in the reference genome; SNPs that potentially introduced new G4 motifs were not analyzed.

The genomic coordinates of 24 243 protein-coding RefSeq genes were downloaded from ftp://ftp.ncbi.nih.gov/refseq (NCBI build 36.3) and used for the annotation of G4 motifs. CpG islands and 28-way vertebrate MultiZ alignments were obtained from the UCSC genome browser (38), available at http://genome.ucsc.edu. Motifs located in four defined genomic regions were subsequently analyzed: 5′ gene regions, 3′ gene regions, the first gene intron and intergenic regions. In order to target regulatory G4 sequences involved in gene transcription, we set the limits of the 5′ region of genes to 2-kb upstream of the transcription start site (TSS) and 1-kb downstream of the TSS. Only non-coding sequences (i.e. UTR) were targeted downstream of the TSS (Figure 1a), since coding sequences exhibit a significant depletion of G4 (39). We are aware that downstream of the TSS, the 5′ region will encompass G4 motifs that could be involved in both transcription and RNA processing. Ideally, one should thus evaluate the upstream and downstream regions of the TSS separately. However, having limited our analysis to the transcriptional aspect of G4, we considered it appropriate to combine the contributions by pre-transcription regulatory G4 (upstream of the TSS) and transcription regulatory G4 (downstream of the TSS). The 3′ end of genes was defined in the same manner as the 5′ end, encompassing 1-kb within 3′ UTR and 2-kb downstream of the transcription stop site. We included an analysis of G4 in the first intron (restricted to the first thousand bp), since this genomic region has shown a particular enrichment of G4 (40). Last, for control purposes, we included G4 motifs located in intergenic regions of the human genome.

Genomic G4 motifs that were found within high-copy repeats (as identified by RepeatMasker and Tandem Repeats Finder) were excluded from the analysis.

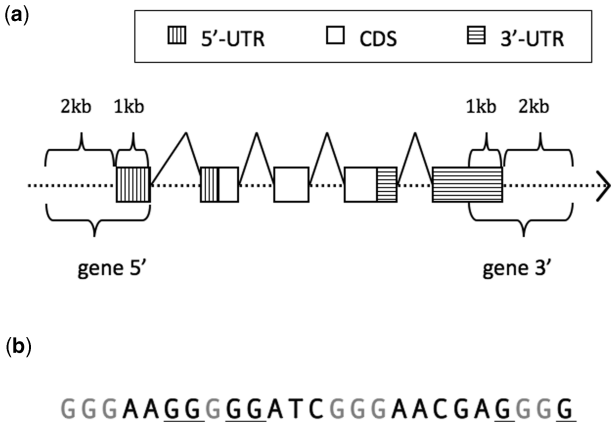

Figure 1. (**a**) A simplified illustration of a human gene, showing how the gene 5′ and gene 3′ regions were defined. (**b**) An example of a G4 sequence motif. The G4-disruptive sites are in grey colour, while the G4-neutral sites are in black. The underlined guanines are guanines within tracts that, when mutated, will not disrupt the G4 consensus.

There were several reasons for this decision. First of all, in the genomic regions of interest (regulatory sequences), the frequency of G4 within unique sequence is nearly twice as that of G4 within repeats. Second, reliable (i.e. validated) SNPs are under-represented in repeats; whereas 51.1% of all reference SNPs in dbSNP are mapped to repeats, only 45.1% of the validated SNPs are located within repeats. Third, in the vertebrate MultiZ alignments, we noted that the availability of reliable alignments for G4 in repeats was poor compared to unique G4.

Non-G4 control sequences

In a search for characteristic patterns of substitutions in the G-rich G4 motifs, we established a set of non-G4 control sequences. The selected non-G4 sequences had the same high GC content as the G4 sequences, but did not match the G4 consensus. This approach enabled us to target differences between G4 and non-G4 unrelated to CpG dinucleotides, since the rate of the most common substitution at CpG dinucleotides (i.e. transition caused by spontaneous hydrolytic deamination of 5-methylcytosine) are dependent on GC content (34,41).

We next provide a short description of the stepwise procedure. For each genomic region analyzed, we created a large library of non-G4 sequence fragments (length 20–28 bp; average length of G4 motifs) that originated either outside or within CpG islands. All fragments were subsequently binned according to GC content. We randomly picked sequence fragments within each bin, the number of fragments being dictated by the probability distribution of G4 motifs with respect to GC content and CpG islands. The SNP density in the total collection of non-G4 fragments was then calculated. This procedure was repeated fifty times for each genomic region and averaged.

RESULTS AND DISCUSSION

Previous studies have demonstrated the importance of computational analyses for the understanding of G4 enrichment in vertebrate genomes (16,17,22,23,28,40, 42–45). In this work, we investigate G4 prevalence from a single nucleotide substitution perspective.

We calculated the density of SNPs in G4 motifs by querying the dbSNP database at the locations of 282 501 motifs in non-repetitive regions of the human genome. Due to the overlapping nature of many G4 motifs (and also some overlapping gene annotations), a number of SNPs were counted more than once in the overall count of SNPs. We checked that this approach did not influence our findings by performing an alternative analysis allowing only one count per SNP in non-overlapping motifs (data not shown). The strandedness of G4 motifs was ignored at this point, and we thus combined the total G4 formation potential involved in either DNA replication or gene transcription.

A total of 10 794 validated SNPs mapped to G4 motifs in the human genome, with an overall density of 1.97 SNPs/kb. With an estimated density of 2.00 SNPs/kb in the genomic background, it was apparent that the level of polymorphism in G4 motifs reflected the genome average. This finding seemed intuitively somewhat unexpected, considering the 2-fold enrichment of hypermutable CpG dinucleotides in G4 compared to the genomic background (Table 1). However, there are two important characteristics of G4 motifs that impose a relatively lower rate of SNPs at CpGs in these sequences. The first feature is the high GC content of G4, since 5-methylcytosine deamination rates are inversely correlated with local GC content (34). Second, there is an extensive overlap between G4 and CpG islands, that is genomic regions in which the cytosines of CpG dinucleotides preferentially remain unmethylated (45,46). Specifically, the coverage density of G4 inside CpG islands was several-fold higher than outside islands (Table 1). The latter observation implies that many G4 CpGs inevitably appear unmethylated in the genome, and this will likely reduce their overall mutagenic potential.

We next sought to identify mutational patterns of G4 motifs that were not related to CpG. To do so, we compared them with a set of randomly picked non-G4 sequences that matched the GC distribution of G4 (see 'Materials and Methods' section). Sampling non-G4 sequences in this manner enabled us to target non-CpG types of pattern in G4, since the mutational characteristics of CpG were approximately equalized between G4 and non-G4. We observed that the SNP density in G4 was consistently lower than in non-G4 sequences, although to a varying extent in the different genomic regions (Figure 2). Since the primary sequence difference between G4 and the random non-G4 fragments was the density of guanine triplets, we hypothesized a suppression of nucleotide polymorphisms in the G4 tetrad regions (i.e. guanine triplets), and that this phenomenon would influence the relative low rate of G4 SNPs.

Critical sites of G4 motifs display low levels of polymorphism

We next investigated whether loop and tetrad (i.e. G-tracts) regions of G4 motifs are subject to different

Table 1. Density of SNPs and CpG dinucleotides in G4 motifs

	Number of G4 motifs	Number of SNPs[a]	SNPs/CpG[b]	CpG island coverage[c]	CpG/kb[d]
Genome	282 501	–	–	–	–
First introns	17 926 (0.33 Mb)	555 (441)	0.00413 (0.014)	0.093 (0.014)	58.4 (28.7)
Gene 5′	31 694 (0.55 Mb)	1157 (874)	0.0052 (0.012)	0.044 (0.010)	57.7 (34.9)
Gene 3′	17 458 (0.30 Mb)	906 (639)	0.0190 (0.038)	0.048 (0.008)	29.5 (13.6)
Intergenic	103 911 (2.01 Mb)	5001 (4096)	0.023 (0.064)	0.036 (0.002)	22.6 (7.6)

[a]Total number of SNPs that map to G4 motifs. The number of unique (non-redundant) SNPs is given in parentheses.
[b]The density estimate of SNPs at G4-CpGs included only C/T and A/G SNPs, since the majority of substitutions occurring at the hypermutable CpG are methylation-dependent transitions. A similar density estimate of SNPs at CpGs in the genomic background is given in parentheses.
[c]Coverage is defined as the fraction of island bases covered by G4 bases. Coverage of G4 outside CpG islands is given in parentheses.
[d]CpG density in genomic background is given in parentheses.

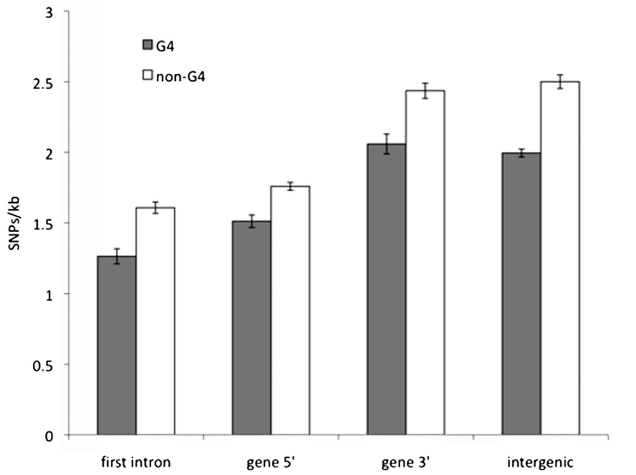

Figure 2. SNP density in G4 sequences versus randomly picked non-G4 sequences. The set of non-G4 sequences were drawn such that their GC-richness was equivalent to that of G4.

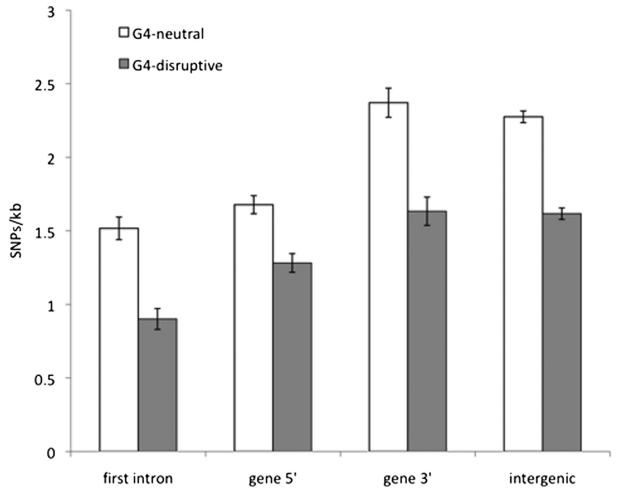

Figure 3. SNP density in G4-disruptive sites versus G4-neutral sites (see Figure 1b for a definition of G4-disruptive and G4-neutral).

mutational pressures. The two distinct G4 regions are important for quadruplex formation and stability, the G-tracts that make up the tetrad planes being critical for formation and folding (32). It is worth noting that, in the G-tracts of G4 motifs, not all substitutions of guanine will disrupt the potential to form a quadruplex structure. For example, if a motif contains a run of four guanines, substitutions at either end of the run will not disrupt the required triplet and could therefore, in principle, preserve the quadruplex-forming potential. On the basis of this reasoning, we classified each position in G4 motifs as either 'G4-disruptive' or 'G4-neutral' (Figure 1b). In all genomic regions analyzed, we found a significantly lower rate of SNPs in G4-disruptive positions relative to the G4-neutral positions (Figure 3). However, since hypermutable CpGs are more frequent at neutral positions than disruptive positions by a factor of nearly three, we performed an additional analysis where CpG sites were masked (Table 2). The difference in SNP density between neutral and disruptive G4 positions decreased when considering non-CpG sites only, though disruptive sites still displayed a significantly lower level of sequence polymorphism. We elaborated on this finding with comparative genomics data, assessing the level of sequence conservation within the two classes of G4 sites. This was accomplished by constructing a four-species multiple sequence alignment (human, monkey, dog and mouse) of G4 motifs from the 28-way vertebrate MultiZ alignments. The disruptive sites of CpG-masked G4 motifs showed consistently higher levels of mammalian sequence conservation than non-disruptive sites (Figure 4).

The evident conserved nature and suppressed level of polymorphisms at G4-disruptive sites could, intuitively, be interpreted as if the G-quadruplex consensus sequence is under functional constraints in the genome. The basic rationale for this argument comes from two recent studies of mutated G-quadruplexes, which demonstrated that their conformational dynamics strongly depends on the position of the mutated guanine (31,32). In an analysis that applied single-molecule FRET spectroscopy on telomeric G4 motifs, the G-quadruplex was severely destabilized when a central guanine was substituted with thymine. Substitutions at the end of a guanine tract also produced less stable structures, though with a far less dramatic effect than the central ones (31). In accordance with these data, we observed a tendency in which the critical guanines of human G4 motifs are less polymorphic than their neutral counterparts. However, we found that this characteristic

Table 2. Density of SNPs in disruptive and neutral sites of G4 sequence motifs

	G4-neutral			G4-disruptive	
	CpGs/kb	SNPs/kb[a]	P[b]	CpGs/kb	SNPs/kb[a]
First introns	166.5	1.52 (1.38)	<0.00001	73.1	0.90 (0.91)
Gene 5′	166.1	1.68 (1.48)	<0.05	71.4	1.28 (1.29)
Gene 3′	83.9	2.37 (1.72)	<0.05	36.1	1.63 (1.45)
Intergenic	65.0	2.28 (1.65)	<0.001	25.6	1.62 (1.46)

[a]Density of SNPs in non-CpG sites are given in parentheses.
[b]Difference in SNP density between G4-disruptive and G4-neutral sites (non-CpG) by Chi-squared analysis.

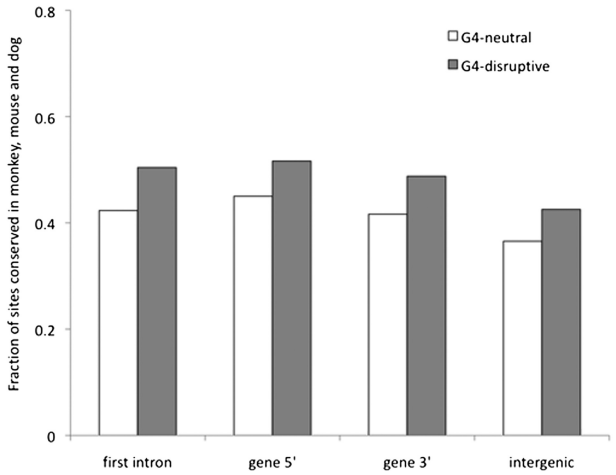

Figure 4. Sequence conservation in G4-disruptive sites versus G4-neutral sites. Shown is the fraction of conserved (i.e. all bases identical) sites at G4-disruptive and G4-neutral sites, as extracted from MultiZ sequence alignments of human G4 with monkey (rheMac2), dog (canFam2) and mouse (mm8). Only non-CpG sites were probed for conservation.

feature of G4 motifs occurred genome-wide, in a strand-independent manner, and also among G4 motifs in intergenic regions. These latter observations suggested that the phenomenon occurs as an effect of intrinsic mutation or DNA repair mechanisms rather than as a consequence of selection for the G4 consensus.

General under-representation of SNPs in guanine tracts

The distribution of SNPs in G4 motifs revealed that nucleotide polymorphisms in G4 DNA would more likely alter the loop conformation than the quadruplex-forming potential. We next asked whether this pattern of guanine substitutions is occurring in a genome-wide fashion, not restricted to the G-tracts of G4 motifs. More specifically, we estimated the relative over-representation of each DNA three-mer at polymorphic guanines by comparing its frequency at polymorphic sites versus non-polymorphic sites, adopting the approach used by Tomso and co-workers (41). For each polymorphic site, two centered three-mers were recorded, one for each allele. Importantly, since the SNP data does not provide any

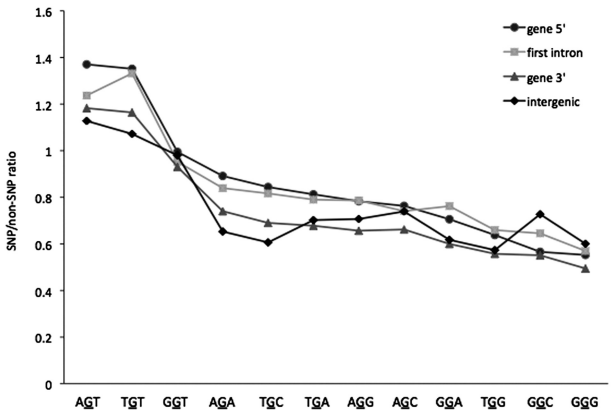

Figure 5. The ratio of DNA three-mers at polymorphic to non-polymorphic sites. Only non-CpG three-mers have been plotted, and each three-mer ratio constitutes the combined ratio of the forward and reverse complementary context. Only SNPs that were proven polymorphic by the HapMap project were used in the calculation.

information as to which strand the original mutational event occurred, we cannot distinguish between a context and its reverse complementary context. We thus ignored strandedness and pooled reverse complementary three-mers together. We confirmed previous observations that CpG-containing three-mers are the most over-represented sequence contexts at human SNPs (36,41). At the opposite end, we observed that a guanine surrounded by other guanines (i.e. 5′-GGG-3′/5′-CCC-3′, polymorphic site underlined), is among the DNA sequence contexts that is most under-represented at polymorphic sites (Figure 5). In fact, it was the most under-represented sequence context among polymorphisms within first introns, at the 5′ end of genes, and at the 3′ end of genes. Our data thus indicate that SNPs with the highest probability of disrupting G-tracts represent the most under-represented SNP context in regulatory gene sequences. We also noted that for sequence contexts at both ends of G-tracts, which for three-mers constitute the 5′-NGG-3′/5′-CCN-3′ and 5′-GGN-3′/5′-NCC-3′ contexts, the frequencies of polymorphisms were generally low. An exception was the 5′-GGT-3′/5′-ACC-3′ context (and the CpG-containing 5′-CGG-3′/5′-CCG-3′, not shown in Figure 5).

Which biological mechanisms could underlie the low rate of polymorphisms inside guanine/cytosine tracts? The phenomenon was not only evident in regulatory regions, but also appeared to occur in intergenic regions, where the modulation of mutational output by natural selection is believed to be weaker. The latter suggests that the observed pattern of SNPs reflects a context-dependency in the mechanisms underlying human mutation. The mutational input to polymorphisms in DNA is considered to be base damage or incorporation of incorrect bases by polymerases during replication, followed by no or error-prone DNA repair. Both the frequency of damages, and the efficiency and fidelity of DNA replication and repair are probably dependent on the sequence context. It is clear that a very significant source of mutations is due to deamination of 5-methylcytosine (5mC) in

CpG dinucleotides. An important additional source of mutations is due to lacking or error-prone repair of 7,8-dihydro-8-oxo-guanine (8-oxoG) in the DNA. It may be caused by UV radiation or oxidative damage to guanine. Several DNA repair systems targets this type of damage, including base excision repair and mismatch repair, but they are not perfect. The damage may occur either to guanines in the nucleotide pool or directly to the guanines in the DNA. In the former case, 8-oxoG may subsequently be incorporated into the DNA unless degraded by the NUDT1 hydrolase (47). If 8-oxoG in the DNA is not removed by the OGG1 glycosylase (48,49), subsequent replication may lead to an adenine being incorrectly incorporated opposite the 8-oxoG instead of a cytosine. If the adenine is not removed by the MUTYH glycosylase (50) before the next round of replication, this process may result in a G:C to T:A transversion. McCulloch *et al.* (51) has recently studied the efficiency and fidelity of DNA in 8-oxoG bypass by polymerases, and their work may indicate a slight dependency on the sequence context for the human polymerase η. Further work is necessary to determine, in detail, the context dependency of polymerases and if this can be a basis for sequence-dependent mutation rates.

Imbalance in the nucleotide precursor pool represents another potential source of mutations. In a mammalian model system that induced thymidine mutations by pool perturbation, it was shown that guanine residues flanked on their 3′ side by other guanine residues are severalfold less mutable than guanine residues flanked on their 3′ side by a different base (52). The underlying mechanism for this pattern was not examined. The authors do, however, argue that differential repair of misincorporated thymidines could be involved. Nonetheless, it is intriguing to see how well these patterns of induced mutations fit with the spectrum we observed for guanine SNPs.

Could systematic DNA-sequencing errors among the polymorphisms collected from dbSNP account for the observed pattern? It has been shown that a few sequence contexts are particularly prone to sequencing errors (one of them being C(A/Y)C), and that these are overrepresented among non-validated SNPs (53). However, our strategy to pick SNPs from dbSNP was designed in a conservative manner (see 'Materials and Methods' section), thereby excluding the majority of false-positive SNPs. Also, we imposed even stricter requirements in the analysis of SNP three-mers, in which we only considered SNPs that were proven polymorphic by HapMap genotyping.

A G4 strand bias for disruptive SNPs

In the previous analyses of SNPs in G4 motifs, we considered general G4 formation potential during DNA denaturation, thereby ignoring the strand orientation of motifs. If we regard G4-regulated gene transcription as a separate process, the potential for regulation lies primarily within motifs on the nontemplate strand, which has shown a significant enrichment relative to the template strand (40,42). We therefore undertook an additional analysis of SNPs in G4 that incorporated strandness of motifs.

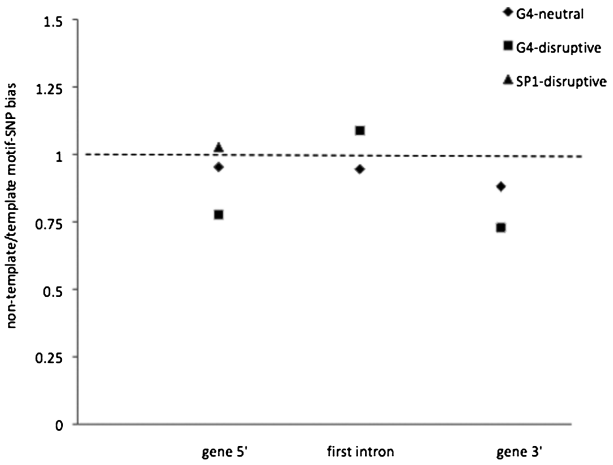

Figure 6. The ratio of SNP density (non-CpG) in nontemplate motifs to the SNP density in template motifs. The dashed line indicates a similar rate of SNPs with respect to the strandedness of the motif, i.e. no strand bias.

The extent of G4 strand bias was defined as the ratio of SNP density (non-CpG) in G4 on the non-template strand to the SNP density in G4 on the template strand, where a ratio of 1 implies no strand bias. Interestingly, we observed a marked strand bias for G4-disruptive SNPs in regulatory sequences, while negligible biases were observed among the neutral G4 SNPs (Figure 6). For disruptive SNPs, it was evident that their density in G4 motifs on the non-template strand was lower than on the template strand. This bias was significant at the 5′ end of genes ($P < 0.02$, $\chi^2 = 5.77$, df = 1) and at the 3′ end of genes ($P < 0.02$, $\chi^2 = 5.82$, df = 1). The result was not an artefact of the overlapping G4 motifs (and SNPs), since the count of unique SNPs in non-overlapping G4 motifs also produced significant strand biases at a significance level of 0.05 (data not shown). As a means to validate the observation at the 5′ end, and to test whether the result was a mere consequence of general suppression of polymorphisms in guanine tracts on the nontemplate strand, we carried out a similar type of analysis with a related sequence element, the SP1 transcription factor (5′-GGGCGG-3′) (44). More specifically, we asked whether there is a strand bias (with respect to SP1) for nucleotide polymorphisms that disrupt the SP1 motif at positions 2 or 3 (two non-CpG sites). The level of SP1 disruption did not differ significantly between the two strands at the 5′ end ($P = 0.855$, $\chi^2 = 0.03$, df = 1), although the set of polymorphisms that mapped to the SP1 motif was considerably smaller than the G4 set (524 SP1 polymorphisms versus 1157 G4 polymorphisms).

The low rate of human SNPs in G4-disruptive positions on the non-template strand support a higher relative importance for this strand in G4-mediated gene regulation. When present on this strand downstream of the TSS, the G-quadruplex may form as part of the pre-mRNA and/or potentially the mRNA, and it may thus serve as multiple targets for regulation (40). The formation of G4 on the template strand would on the other hand hinder the progression of the RNA polymerase, and is

therefore less desirable (23). We also showed that another G-rich element, the SP1 transcription factor, did not display any strand bias with respect to disruptive SNPs at the 5′ end. It may thus seem as if the pattern supports a specific biological importance for G4 motifs on the non-template strand at the 5′ end of genes.

CONCLUSION

The recent genome-wide scans of G4 motifs in the human genome have identified enrichment in gene regulatory sequences, and the same tendency has been shown when searching the genomes of chimpanzee, rat and mouse (16,17,45). The prevalence of G4 motifs upstream of mammalian genes has been interpreted as a sign of selection for G4, and consequently implicated the G-quadruplex structure as a potential mechanism for regulating gene expression (23,39). On the basis of sequence data only, it is nonetheless impossible to determine the extent of quadruplex formation *in vivo*, although it seems most likely that only a low percentage of the G4 motifs will adopt structures during denaturation.

Here, a close examination of the context-dependent pattern of guanine polymorphisms has provided an additional perspective on G4 prevalence. It shows how the aspect of sequence mutagenesis could impact the evolution of guanine tracts, the key component in G4 motifs. Although significant patterns emerged, our results are limited by the approximately 11 000 SNPs that map to G4 motifs in the human genome. Following next-generation sequencing and collaborative efforts such as the 1000 Genome Projects (54), more data should be available for studying the nature of G4 sequence polymorphism. An interesting extension of our analysis, which requires more validated SNPs available, is to relate the directionality of each SNP (i.e. by determining the ancestral and derived allele) to G4 evolution. Nevertheless, in light of our current findings, we warrant a closer examination of the relationship between G4 and other factors that might constrain the nearest-neighbour sequence patterns in DNA, an example being the physical requirements needed for the dense packing of DNA around nucleosomes.

FUNDING

Research Council of Norway. Funding for open access charge: the EU FP7 contract 223367.

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. Hazel,P., Huppert,J., Balasubramanian,S. and Neidle,S. (2004) Loop-length-dependent folding of G-quadruplexes. *J. Am. Chem. Soc.*, **126**, 16405–16415.
4. Risitano,A. and Fox,K.R. (2004) Influence of loop size on the stability of intramolecular DNA quadruplexes. *Nucleic Acids Res.*, **32**, 2598–2606.
5. Burge,S., Hazel,P. and Todd,A.K. (2006) Quadruplex DNA: sequence, topology and structure. *Nucleic Acids Res.*, **34**, 5402–5415.
6. 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.
7. Sundquist,W.I. and Klug,A. (1989) Telomeric DNA dimerizes by formation of guanine tetrads between hairpin loops. *Nature*, **342**, 825–829.
8. Williamson,J.R., Raghuraman,M.K. and Cech,T.R. (1989) Monovalent cation-induced structure of telomeric DNA: the G-quartet model. *Cell*, **59**, 871–880.
9. Schaffitzel,C., Berger,I., Postberg,J., Hanes,J., Lipps,H.J. and Pluckthun,A. (2001) In vitro generated antibodies specific for telomeric guanine-quadruplex DNA react with Stylonychia lemnae macronuclei. *Proc. Natl Acad. Sci. USA*, **98**, 8572–8577.
10. 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.
11. 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*. *Nat. Struct. Mol. Biol.*, **12**, 847–854.
12. Bachrati,C.Z. and Hickson,I.D. (2006) Analysis of the DNA unwinding activity of RecQ family helicases. *Methods Enzymol.*, **409**, 86–100.
13. 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.
14. Wu,Y., Shin-ya,K. and Brosh,R.M. Jr. (2008) FANCJ helicase defective in Fanconia anemia and breast cancer unwinds G-quadruplex DNA to defend genomic stability. *Mol. Cell Biol.*, **28**, 4116–4128.
15. Fry,M. (2007) Tetraplex DNA and its interacting proteins. *Front. Biosci.*, **12**, 4336–4351.
16. Huppert,J.L. and Balasubramanian,S. (2005) Prevalence of quadruplexes in the human genome. *Nucleic Acids Res.*, **33**, 2908–2916.
17. Todd,A.K., Johnston,M. and Neidle,S. (2005) Highly prevalent putative quadruplex sequence motifs in human DNA. *Nucleic Acids Res.*, **33**, 2901–2907.
18. Kikin,O., D'Antonio,L. and Bagga,P.S. (2006) QGRS Mapper: a web-based server for predicting G-quadruplexes in nucleotide sequences. *Nucleic Acids Res.*, **34**, W676–W682.
19. 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.
20. 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.
21. Wang,Y. and Patel,D.J. (1993) Solution structure of the human telomeric repeat d[AG3(T2AG3)3] G-tetraplex. *Structure*, **1**, 263–282.
22. Huppert,J.L. and Balasubramanian,S. (2007) G-quadruplexes in promoters throughout the human genome. *Nucleic Acids Res.*, **35**, 406–413.
23. Du,Z., Zhao,Y. and Li,N. (2008) Genome-wide analysis reveals regulatory role of G4 DNA in gene transcription. *Genome Res.*, **18**, 233–241.
24. 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.
25. Simonsson,T., Pecinka,P. and Kubista,M. (1998) DNA tetraplex formation in the control region of c-myc. *Nucleic Acids Res.*, **26**, 1167–1172.
26. Fernando,H., Reszka,A.P., Huppert,J., Ladame,S., Rankin,S., Venkitaraman,A.R., Neidle,S. and Balasubramanian,S. (2006)

A conserved quadruplex motif located in a transcription activation site of the human c-kit oncogene. *Biochemistry*, **45**, 7854–7860.
27. Yafe,A., Etzioni,S., Weisman-Shomer,P. and Fry,M. (2005) Formation and properties of hairpin and tetraplex structures of guanine-rich regulatory sequences of muscle-specific genes. *Nucleic Acids Res.*, **33**, 2887–2900.
28. 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.
29. Bugaut,A. and Balasubramanian,S. (2008) A sequence-independent study of the influence of short loop lengths on the stability and topology of intramolecular DNA G-quadruplexes. *Biochemistry*, **47**, 689–697.
30. Kumar,N., Sahoo,B., Varun,K.A. and Maiti,S. (2008) Effect of loop length variation on quadruplex-Watson Crick duplex competition. *Nucleic Acids Res.*, **36**, 4433–4442.
31. Lee,J.Y. and Kim,D.S. (2009) Dramatic effect of single-base mutation on the conformational dynamics of human telomeric G-quadruplex. *Nucleic Acids Res.*, **37**, 3625–3634.
32. Gros,J., Rosu,F., Amrane,S., De Cian,A., Gabelica,V., Lacroix,L. and Mergny,J.L. (2007) Guanines are a quartet's best friend: impact of base substitutions on the kinetics and stability of tetramolecular quadruplexes. *Nucleic Acids Res.*, **35**, 3064–3075.
33. Blake,R.D., Hess,S.T. and Nicholson-Tuell,J. (1992) The influence of nearest neighbors on the rate and pattern of spontaneous point mutations. *J. Mol. Evol.*, **34**, 189–200.
34. Fryxell,K.J. and Moon,W.J. (2005) CpG mutation rates in the human genome are highly dependent on local GC content. *Mol. Biol. Evol.*, **22**, 650–658.
35. Hodgkinson,A., Ladoukakis,E. and Eyre-Walker,A. (2009) Cryptic variation in the human mutation rate. *PLoS Biol.*, **7**, e27.
36. Krawczak,M., Ball,E.V. and Cooper,D.N. (1998) Neighboring-nucleotide effects on the rates of germ-line single-base-pair substitution in human genes. *Am. J. Hum. Genet.*, **63**, 474–488.
37. Sherry,S.T., Ward,M.H., Kholodov,M., Baker,J., Phan,L., Smigielski,E.M. and Sirotkin,K. (2001) dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.*, **29**, 308–311.
38. Karolchik,D., Baertsch,R., Diekhans,M., Furey,T.S., Hinrichs,A., Lu,Y.T., Roskin,K.M., Schwartz,M., Sugnet,C.W., Thomas,D.J. *et al.* (2003) The UCSC Genome Browser Database. *Nucleic Acids Res.*, **31**, 51–54.
39. Eddy,J. and Maizels,N. (2009) Selection for the G4 DNA motif at the 5′ end of human genes. *Mol. Carcinog.*, **48**, 319–325.
40. Eddy,J. (2007) Conserved elements with potential to form polymorphic G-quadruplex structures in the first intron of human genes. *Nucleic Acids Res.*, **36**, 1321–1333.
41. Tomso,D.J. and Bell,D.A. (2003) Sequence context at human single nucleotide polymorphisms: overrepresentation of CpG dinucleotide at polymorphic sites and suppression of variation in CpG islands. *J. Mol. Biol.*, **327**, 303–308.
42. 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.
43. Huppert,J.L., Bugaut,A., Kumari,S. and Balasubramanian,S. (2008) G-quadruplexes: the beginning and end of UTRs. *Nucleic Acids Res.*, **36**, 6260–6268.
44. Todd,A.K. and Neidle,S. (2008) The relationship of potential G-quadruplex sequences in cis-upstream regions of the human genome to SP1-binding elements. *Nucleic Acids Res.*, **36**, 2700–2704.
45. Verma,A., Halder,K., Halder,R., Yadav,V.K., Rawal,P., Thakur,R.K., Mohd,F., Sharma,A. and Chowdhury,S. (2008) Genome-wide computational and expression analyses reveal G-quadruplex DNA motifs as conserved cis-regulatory elements in human and related species. *J. Med. Chem.*, **51**, 5641–5649.
46. Bird,A.P. (1986) CpG-rich islands and the function of DNA methylation. *Nature*, **321**, 209–213.
47. Sakumi,K., Furuichi,M., Tsuzuki,T., Kakuma,T., Kawabata,S., Maki,H. and Sekiguchi,M. (1993) Cloning and expression of cDNA for a human enzyme that hydrolyzes 8-oxo-dGTP, a mutagenic substrate for DNA synthesis. *J. Biol. Chem.*, **268**, 23524–23530.
48. Bjoras,M., Luna,L., Johnsen,B., Hoff,E., Haug,T., Rognes,T. and Seeberg,E. (1997) Opposite base-dependent reactions of a human base excision repair enzyme on DNA containing 7,8-dihydro-8-oxoguanine and abasic sites. *EMBO J.*, **16**, 6314–6322.
49. Nash,H.M., Bruner,S.D., Scharer,O.D., Kawate,T., Addona,T.A., Spooner,E., Lane,W.S. and Verdine,G.L. (1996) Cloning of a yeast 8-oxoguanine DNA glycosylase reveals the existence of a base-excision DNA-repair protein superfamily. *Curr. Biol.*, **6**, 968–980.
50. Slupska,M.M., Baikalov,C., Luther,W.M., Chiang,J.H., Wei,Y.F. and Miller,J.H. (1996) Cloning and sequencing a human homolog (hMYH) of the Escherichia coli mutY gene whose function is required for the repair of oxidative DNA damage. *J. Bacteriol.*, **178**, 3885–3892.
51. McCulloch,S.D., Kokoska,R.J., Garg,P., Burgers,P.M. and Kunkel,T.A. (2009) The efficiency and fidelity of 8-oxo-guanine bypass by DNA polymerases {delta} and {eta}. *Nucleic Acids Res.*, **37**, 2830–2840.
52. Kresnak,M.T. and Davidson,R.L. (1992) Thymidine-induced mutations in mammalian cells: sequence specificity and implications for mutagenesis in vivo. *Proc. Natl Acad. Sci. USA*, **89**, 2829–2833.
53. Platzer,M., Hiller,M., Szafranski,K., Jahn,N., Hampe,J., Schreiber,S., Backofen,R. and Huse,K. (2006) Sequencing errors or SNPs at splice-acceptor guanines in dbSNP? *Nat. Biotechnol.*, **24**, 1068–1070.
54. Siva,N. (2008) 1000 Genomes project. *Nat. Biotechnol.*, **26**, 256.