Published online 3 March 2011 *Nucleic Acids Research, 2011, Vol. 39, No. 12* ***4975–4983***
doi:10.1093/nar/gkr079

G4 motifs correlate with promoter-proximal transcriptional pausing in human genes

Johanna Eddy[1,†], Aarthy C. Vallur[2], Sudir Varma[3], Hongfang Liu[3,4], William C. Reinhold[3], Yves Pommier[3] and Nancy Maizels[1,2,5,*]

[1]Molecular and Cellular Biology Graduate Program, [2]Department of Immunology, [3]Laboratory of Molecular Pharmacology, Center for Cancer Research, National Cancer Institute, NIH, Bethesda, MD 20892, [4]Lombardi Comprehensive Cancer Center, Georgetown University Medical Center and [5]Department of Biochemistry, University of Washington School of Medicine, Seattle, WA 98195-7650, USA

Received December 19, 2010; Revised January 26, 2011; Accepted January 31, 2011

ABSTRACT

The RNA Pol II transcription complex pauses just downstream of the promoter in a significant fraction of human genes. The local features of genomic structure that contribute to pausing have not been defined. Here, we show that genes that pause are more G-rich within the region flanking the transcription start site (TSS) than RefSeq genes or non-paused genes. We show that enrichment of binding motifs for common transcription factors, such as SP1, may account for G-richness upstream but not downstream of the TSS. We further show that pausing correlates with the presence of a GrIn1 element, an element bearing one or more G4 motifs at the 5′-end of the first intron, on the non-template DNA strand. These results suggest potential roles for dynamic G4 DNA and G4 RNA structures in *cis*-regulation of pausing, and thus genome-wide regulation of gene expression, in human cells.

INTRODUCTION

Genome-wide studies have shown that Pol II transcription complexes pause just downstream of the transcription start site (TSS) at many human genes (1–3). Pausing may poise a polymerase for rapid induction of transcription upon receipt of the appropriate signal, or provide a checkpoint at which the transcription complex ensures that all factors are present for productive elongation. Pausing occurs only at a fraction of genes, so one or more features of genomic sequence or structure must contribute to pausing at human genes. Those features have not yet been defined. Identification of the local features of DNA architecture that contribute to DNA pausing has important implications for understanding mechanisms of genomic instability and the response of cells to chemotherapeutics.

G-rich intron 1 (GrIn1) elements are a recently identified feature of genomic structure (4). These conserved elements are present in almost one-half of all human genes and map to the 5′-end of the first intron and the non-template strand. They bear the signature sequence motif characteristic of regions with potential to form G4 structures, $G_{\geq 3}N_xG_{\geq 3}N_xG_{\geq 3}N_xG_{\geq 3}$ (5–8). Their G-richness cannot be accounted for by sequences that would make them targets of well-defined regulatory mechanisms, such as CpG dinucleotides that undergo methylation, or motifs recognized by transcription factors or RNA processing factors. GrIn1 elements occupy a privileged genomic position, as they are located on average 200 nt downstream of the TSS, within 100 bp of the 5′-end of the first intron and on the non-template strand. An element at this intronic position may regulate transcription or RNA processing without conferring selective pressure on protein sequence.

The position, conservation and abundance of GrIn1 elements suggest that these elements might function in regulation of gene expression. The G-richness of the GrIn1 element confers the potential to form a dynamic structure upon transcription of a genomic region. This structure, called a G-loop, carries a co-transcriptional RNA/DNA hybrid on the template strand, and G4 DNA interspersed with single-stranded regions on the G-rich non-template (coding) strand (9–12). Persistent co-transcriptional RNA/DNA hybrids like those that

*To whom correspondence should be addressed. Tel: +206 221 6876; Fax: +206 221 6781; Email: maizels@u.washington.edu
Reprint Requests to Nancy Maizels, Department of Immunology and Department of Biochemistry, University of Washington School of Medicine, Box 357650, Seattle, WA 98195-7650. Tel: +206 221 6876; Fax: +206 221 6781; Email: maizels@u.washington.edu
†This article is dedicated to the memory of Johanna Eddy who passed away soon after receiving her PhD degree, following a courageous struggle with breast cancer. Her creativity and eagerness to rise to every challenge will continue to inspire everyone lucky enough to have worked with her.

characterize G-loops can contribute to genomic instability (11,13–16). They also prolong the denaturation of the DNA strands that normally accompanies transcription, enhancing the potential of DNA to form G4 structures that may function as regulatory targets.

Here, we address the possibility that GrIn1 elements correlate with transcriptional pausing. We show that genes that can be classified as paused are more G-rich in the region flanking the TSS than RefSeq genes or non-paused genes, and we demonstrate that there is a strong correlation between transcriptional pausing and the presence of a GrIn1 element. These results suggest that formation of G4 structures on the non-template strand of the DNA or at the 5′-end of the nascent mRNA may promote promoter proximal pausing. GrIn1 elements may thereby contribute to genome-wide regulation of gene expression of specific classes of genes and they may also influence cellular sensitivity to drugs that perturb the normal dynamics of formation of DNA structure during transcription, including topoisomerase poisons and compounds designed to target G4 structures.

MATERIALS AND METHODS

Sequence data, regulatory motif masking and statistical analysis

Sequence data for the 18 187 human RefSeq genes (NCBI 36 assembly) were downloaded from the Ensembl database 54 using BioMart (17,18). As previously, we defined G-richness as the frequency within each set of genes of 100 nt sequence that contains a G4 DNA signature motif, $G_{\geq 3}N_xG_{\geq 3}N_xG_{\geq 3}N_xG_{\geq 3}$ (5). Intron sequence derivation and calculations of G-richness were performed as described (4). For genes that express alternative transcripts with different first introns, the 5′-most first intron was included in the analysis. Masking of regulatory motifs was performed as described (4). The χ^2-test was performed with the statistics program R version 2.7.1.

Microarray analysis of NCI-60 lines

Affymetrix GeneChip Human Exon 1.0 ST (GH Exon 1.0 ST) microarray analysis of NCI-60 cancer cell lines was carried out as described previously (8). In brief, microarrays were hybridized, usually in triplicate, following manufacturer's instructions at GeneLogic (Gaithersburg, MD), and results normalized by robust multi-array analysis (19) using Partek Genomics Suite version 6.3. The GH Exon 1.0 ST microarray analysis of the NCI-60 lines characterized expression of 16 959 annotated genes, and probes were mapped to transcripts using exon designations assigned by SpliceCenter (20). Classification of genes as paused or non-paused was based on the difference in average probe set intensity level of expression and average standard deviation between the first exon and the other exons across all cell lines. Probe intensity criteria were first developed empirically for the topoisomerase 1 (TOP1) gene (8), and those criteria were applied to define paused genes from the larger database. At TOP1, it was noted that exon 1 was expressed both at higher level and in a manner less variable than the other exons. The increase in expression level of exon 1 was 1.24, and the reduction in standard deviation was 0.244. For subsequent analyses, paused genes were defined as exhibiting a difference in average intensities of the exons 2 through N as compared to exon 1 less than −1.24, and an increase in the standard deviations of the exons 2 through N as compared with exon 1 of greater than 0.244. Genes were classified as non-paused if they exhibited a difference in average intensities of the exons 2 through N as compared to exon 1 ≥0, and a decrease in average standard deviation of less than zero. Using these criteria, 3165 (19%) genes were classified as paused and 1401 (8%) as non-paused. The remaining genes did not fall into either category.

RESULTS

Inverse correlation between Pol II binding and G-richness

In human genes, two peaks of G-richness flank the TSS, centred on the region −100 to +1 and +200 to +300 (4). To ask if these peaks of G-richness correlate with binding by RNA Pol II, we graphed the frequency of G-richness and of Pol II binding sites as determined by Chromatin Immunoprecipitation-Sequencing (ChIP-Seq) for human T cells (1) in the 2 kb region flanking the TSS. The peak of Pol II binding, near +100, corresponded to a local minimum of G-richness, 200 bp downstream from the peak, near +300 (Figure 1). This peak represents the average of all Pol II molecules, regardless of pausing status. That the peak of Pol II binding coincides with a local minimum of G-richness is consistent with the A/T richness of most promoters.

CpG dinucleotides, which are sites for regulatory methylation, can contribute to local G-richness. We graphed the distribution of CpG dinucleotides in the region flanking the TSS, and showed that this comprised a relatively broad peak, which is not coincident with the peaks of G-richness and lies somewhat upstream of the peak of Pol II binding (Figure 1).

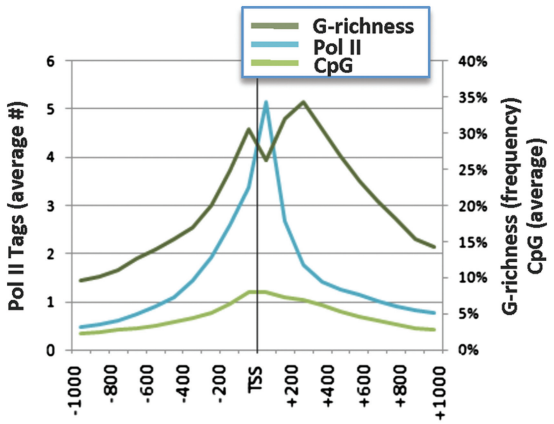

Figure 1. Inverse correlation between Pol II binding and G-richness at the TSS. Graph of the frequency Pol II binding sites (1), CpG dinucleotides, and the frequency of G-richness in the interval −1000 to +1000 around the TSS. G-richness was defined as the frequency within each set of genes of 100 nt sequences containing the G4 DNA signature motif, $G_{\geq 3}N_xG_{\geq 3}N_xG_{\geq 3}N_xG_{\geq 3}$ (4).

Promoters of paused genes are enriched in G4 motifs

We next asked whether G-richness correlates with pausing, using three different operational definitions to classify genes as paused. One of these definitions distinguishes paused and non-paused genes based on relative expression of exon 1 and downstream exons, as determined by microarray analysis. The NCI-60 panel of cell lines includes 60 cell lines representing multiple tumor types for which drug sensitivity and transcriptome activity have been extensively studied and correlated (8,21–23). We calculated the frequency of G-richness in the region −1000 to +1000 for the genes classified as paused (19%) or not paused (8%) across all cell lines in the NCI-60 panel database, and for all RefSeq genes. Paused genes were more G-rich than RefSeq genes or than non-paused genes (Figure 2A).

A second operational definition identifies paused genes as those at which Pol II is stably associated with the TSS in the absence of gene expression. Approximately one-third of genes in primary resting human $CD4^+$ T cells were classified as paused by this criterion (1). We calculated the frequency of G-richness of the region flanking the TSS of those genes relative to RefSeq genes and non-paused genes in that data set. This analysis showed that paused genes were more G-rich in the region flanking the TSS than other genes (Figure 2B).

Chromatin marks can also be used to distinguish paused and non-paused genes. Histone modifications correlate with gene expression, with H3K4me3 characterizing active genes and H3K27me3 characterizing repressed genes. Some genes carry bivalent chromatin marks, with H3K4me3 near the promoter and H3K27me3 distributed more broadly along the gene and such bivalent marks can be used to distinguish paused genes from other genes (1,24). Calculation of the frequency of G-richness in the region from −1000 to +1000 showed that genes with bivalent H3K4me3 and H3K27me3 marks were more G-rich than RefSeq genes or inactive genes with monovalent H3K27me3 marks (Figure 2C).

The above analyses show that paused genes as defined by any of the three above criteria are more G-rich than non-paused genes. The G-richness of paused genes extends throughout the 2 kb interval analyzed, and includes regions both upstream and downstream of the promoter. Sequences upstream of the promoter may contribute to pausing by serving as sites for transcription factors that promote pausing. In this regard, it is interesting that genes classified as non-paused based on relative expression of exon 1 and downstream exons were comparatively G-poor (Figure 2A). This raises the possibility that transcription factors with G/C rich binding motifs may contribute to pausing at some genes, or conversely that transcription factors with A/T rich binding motifs may prevent pausing at others.

Strand biased G-richness downstream of the TSS at paused genes

The results above (Figure 2) establish that paused genes are more G-rich than other genes. How might G-richness contribute to the mechanism of pausing? Pausing could in

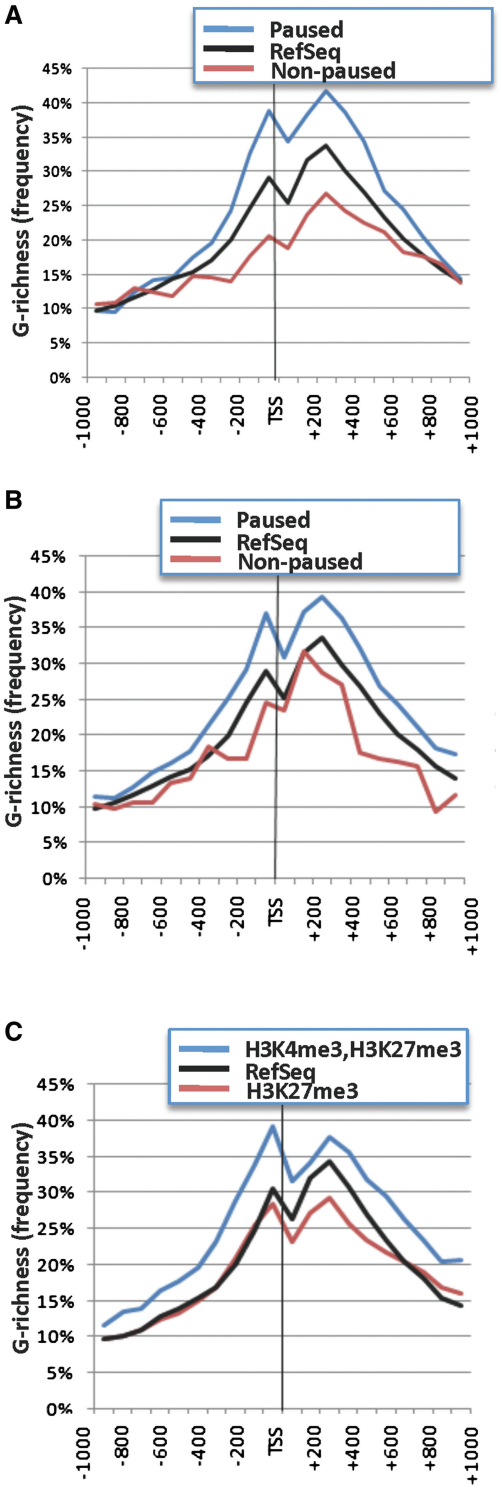

Figure 2. G4 motifs are enriched near promoters of paused genes. (**A**) Graph of the frequency of G-richness in genes defined as paused from the NCI-60 database in the interval −1000 to +1000 around the TSS. (**B**) Graph of the frequency of G-richness in genes in which Pol II is stably associated with the TSS in the absence of gene expression in the interval −1000 to +1000 around the TSS. This data set derives from analysis of primary resting human $CD4^+$ T cells (1), and corresponds to the same data set for which genome-wide analysis Pol II position is presented in Figure 1. (**C**) Graph of the frequency of G-richness in the interval −1000 to +1000 around the TSS in genes carrying bivalent chromatin marks H3K4me3 and H3K27me3, as determined by analysis of primary resting human $CD4^+$ T cells (1).

principle be caused by formation of G4 structures in either the DNA or the nascent transcript. If G-loop formation contributes to the mechanism of pausing, then G-richness of paused genes is predicted to exhibit a strand bias, with G-rich regions downstream of the TSS concentrated in the non-template strand (9,10,12). We therefore compared the frequency of non-template and template strand G-richness in the 2 kb region spanning the TSS for genes classified as paused and non-paused based on relative expression of exon 1 and downstream exons in the NCI-60 database, and for all RefSeq genes. For all three groups of genes, there was clear strand asymmetry in G-richness downstream of the TSS, with greater G-richness on the non-template strand. Notably, paused genes were more G-rich than RefSeq genes, which were more G-rich than non-paused genes (Figure 3).

G-richness of the genes analyzed, exhibited a characteristic distribution. For all three groups, more genes were G-rich on the non-template strand than on the template strand.

For all three groups of genes, upstream of the TSS and on the non-template strand, the maximum frequency of G-richness fell within the region from −100 to −1, where 40% of paused genes were G-rich, compared with 22% of non-paused genes and 30% of all RefSeq genes. Downstream of the TSS and on the non-template strand, maximum frequency of G-richness fell within the region from +200 to +300, where 42% of the paused genes were G-rich, compared with 28% of the non-paused genes and 35% of the RefSeq genes. Downstream of the TSS and on the non-template strand, a peak in the frequency of G-richness was also evident among paused and RefSeq genes, but not non-paused genes.

Transcriptional regulatory motifs account for some but not all G-richness near the TSS of paused genes

G-richness can reflect the presence of DNA sequence elements with well-characterized functions, including CpG dinucleotides that are targets of methylation as well as motifs for some common transcription factors that recognize G-rich sites in duplex DNA, including SP1 (RGGCGKR), KLF (GGGGTGGGG), EKLF (AGGG TGKGG), MAZ (GGGAGGG), EGR-1 (GCGTGGGC G) and AP-2 (CGCCNGSGGG). To eliminate contributions from these elements, we analyzed the distribution of G-richness with these sites masked. The frequency of G-richness may be greatly underestimated following masking, because masking is carried out based on DNA sequence alone, independent of information on whether a motif actually serves as a binding site for its cognate factor. Moreover, in the absence of knowledge regarding whether a specific motif contributes to pausing, masking may even eliminate from the tally genes bearing a motif that promotes pausing. Nonetheless, masking provides a convenient view of how canonical motifs affect the genomic landscape.

We first masked SP1 binding motifs, separately analyzing all RefSeq genes and the paused and non-paused genes identified in the NCI-60 database. Eliminating SP1 motifs primarily affected the region upstream of the TSS, eliminating the peak of G-richness upstream of the TSS in the non-template (but not template) strand of all three sets of genes (Figure 4A). Downstream of the TSS, G-richness of the non-template strand for paused genes was still greater (37%) than for non-paused genes (24%) or RefSeq genes (31%).

Masking of CpG dinucleotides reduced the frequency of G-richness both upstream and downstream of the TSS in all three sets of genes (Figure 4B). Even after masking, there was clear strand asymmetry in G-richness downstream of the TSS for all groups of genes. In addition, G-richness of paused genes remained greater at both upstream and downstream peaks (25 and 24%, respectively) than G-richness of non-paused genes (15 and 20%) or RefSeq genes (19 and 22%). Thus, although CpG content corresponded with pausing in both upstream and downstream regions, it did not account for all of the G-richness surrounding the TSS. We note that a peak of G-richness upstream of the TSS that was eliminated by masking SP1 motifs (Figure 4A) persisted after masking CpG motifs (Figure 4B),

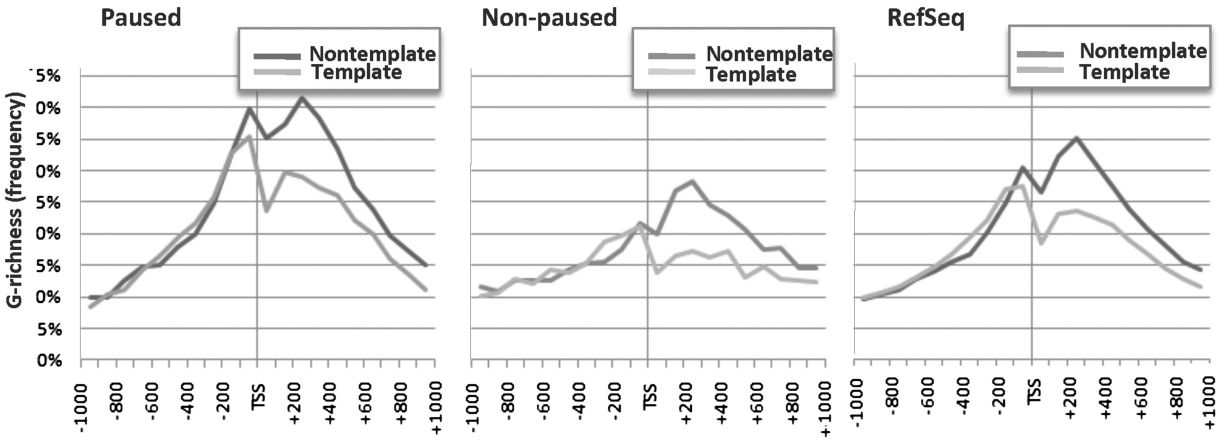

Figure 3. Strand-biased G-richness downstream of the TSS at paused genes. Graph of the frequency of G-richness of the non-template (dark lines) and template (pale lines) strands in the interval −1000 to +1000 around the TSS for genes in the NCI-60 database classified as paused (left) or non-paused (center), and RefSeq genes (right).

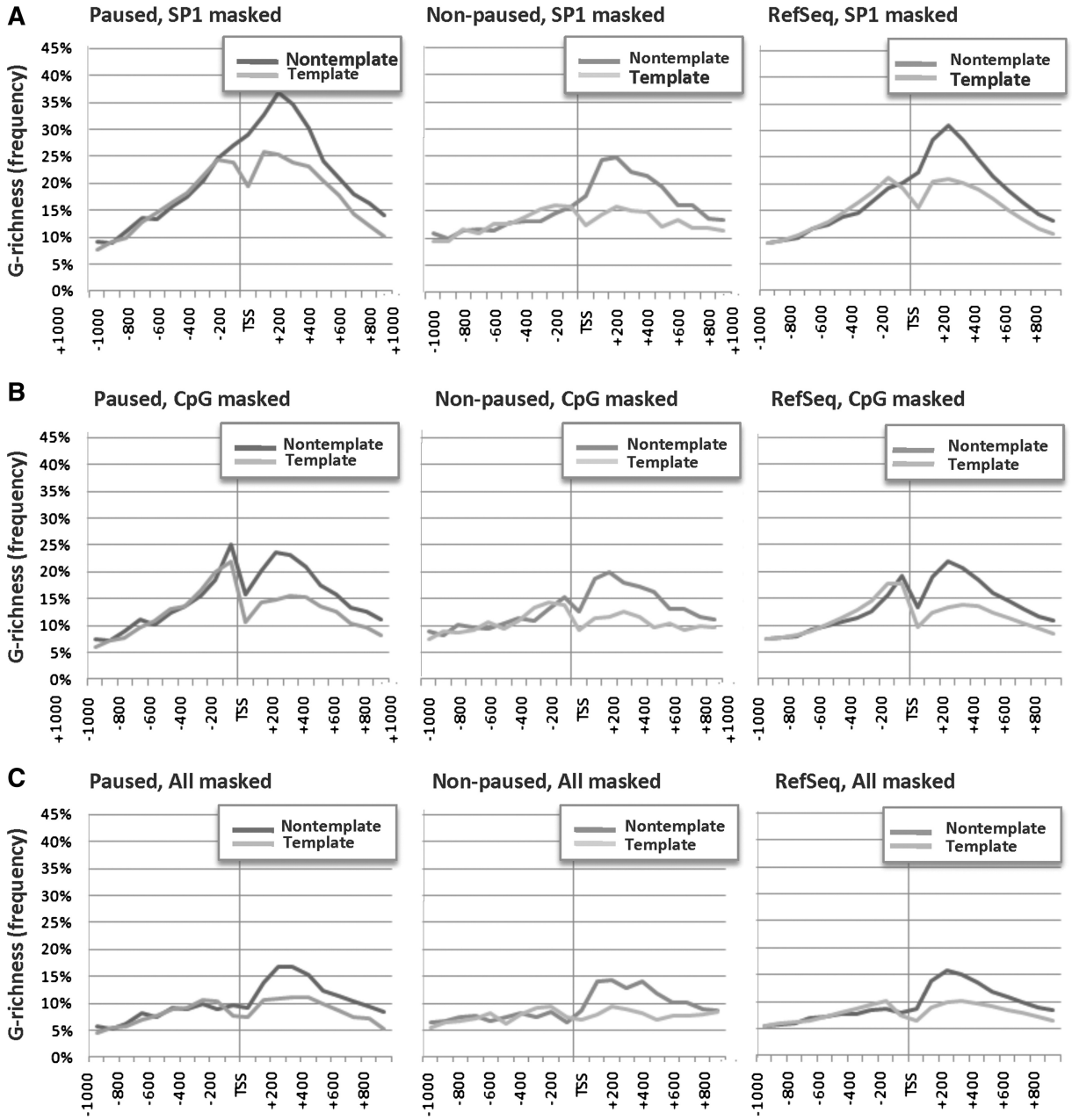

Figure 4. Transcriptional regulatory motifs do not account for G-richness near TSS of paused genes. Graph of the frequency of G-richness of non-template (dark lines) and template (pale lines) strands in the interval −1000 to +1000 around the TSS for paused genes (left), non-paused genes (centre) and all human RefSeq genes (right), with the following motifs masked: (**A**) SP1 motifs. (**B**) CpG motifs. (**C**) CpG, SP1, MAZ, KLF, EKLF, EGR-1 and AP-2 motifs.

suggesting that SP1 motifs make their primary contribution to G-richness upstream and not downstream of the TSS.

Finally, we maximally depleted common G-rich motifs by masking binding motifs for six common transcription factors that bind G-rich sites, including SP1, KLF, EKLF MAZ, EGR-1 and AP-2, as well as CpG motifs. To maximize depletion of these canonical motifs, they were masked before eliminating CpG motifs. This stringent masking diminished the peaks of G-richness upstream and downstream of the TSS in all three classes of genes, but affected the upstream peak most profoundly (Figure 4C). Following stringent masking, the strand asymmetry in G-richness downstream of the TSS persisted, although only small differences were evident in non-template strand G-richness at both upstream and downstream peaks of paused genes (11 and 17%, respectively) relative to non-paused genes (7 and 14%, respectively). The very high stringency of masking is likely to be responsible for this considerable decrease in frequency of G-richness, and these small differences are unlikely to be significant.

GrIn1 elements correlate with pausing

We previously found that almost one-half of all human genes contain G-rich elements on the non-template DNA strand at the 5′-end of the first intron, referred to as GrIn1 elements (4). To ask if a difference in GrIn1 element frequency characterizes paused and non-paused genes, we calculated G-richness for 1000 bp of the first introns for RefSeq genes and genes classified as paused or non-paused in the NCI-60 database. (This analysis was restricted to introns at least 1000 bp in length in order to include a constant number of genes along the length distribution. We previously showed (4) that setting the lower limit of intron size to either 100 bp or 1000 bp generates an essentially identical distribution of G-richness.) A fraction of genes in all three groups exhibited a peak of G-richness at the very 5′-end of the first intron, consistent with the presence of a GrIn1 element (Figure 5A). GrIn1 elements were present in 57% of paused genes, 38% of non-paused genes and 50% of RefSeq genes. The difference between the fraction of paused and non-paused genes containing GrIn1 elements was highly significant ($\chi^2 = 82$; $P < 10^{-10}$).

Motifs for hnRNP proteins and CpG dinucleotides contribute to but do not account for GrIn1 elements

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 contribute to but are not sufficient for binding, so the tally of motifs will overestimate their functional contribution of to G-richness of the intron. After masking these motifs, 34% of paused genes, 19% of non-paused genes and 28% of RefSeq genes retained a peak of G-richness, differences comparable with those observed upon analyzing the unmasked genes (Figure 5B). Masking CpG motifs in addition to hnRNP A and H binding motifs reduced the frequency of G-richness at the 5′-end of intron 1, so that a peak of G-richness was evident in 19% of paused genes, 13% of non-paused genes and 17% of RefSeq genes (Figure 5C). Thus, even with all these motifs masked, there was a greater frequency of GrIn1 elements in paused genes than in other gene classes.

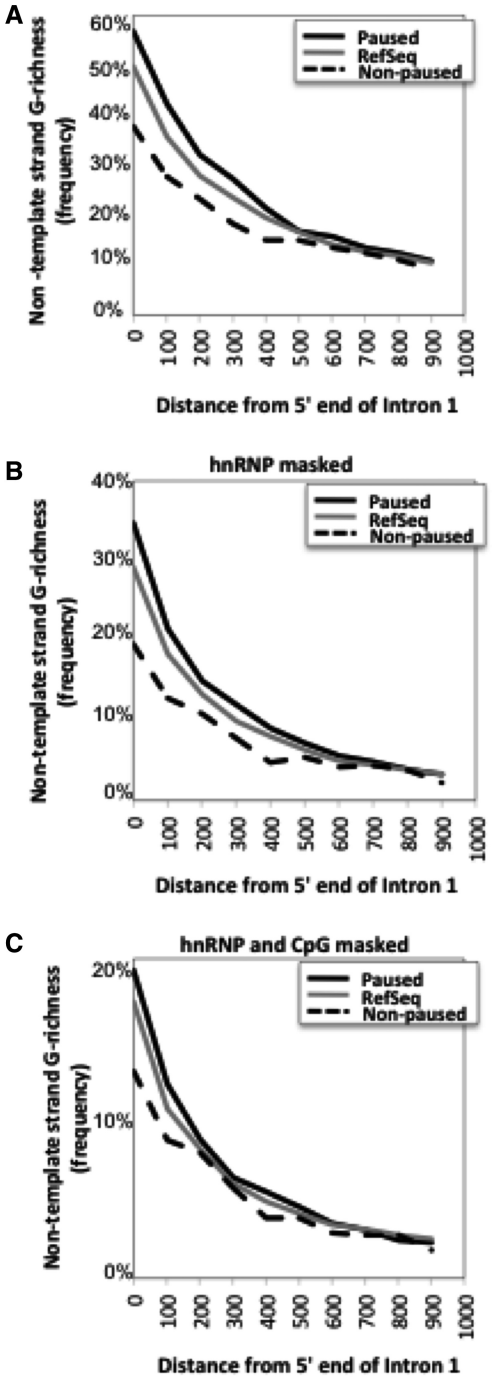

Figure 5. GrIn1 elements correlate with pausing. (**A**) Graph of the frequency of non-template strand G-richness within 1 kb downstream of the TSS for paused, non-paused genes and RefSeq genes (top). (**B**) Graph of the frequency of non-template strand G-richness as in Panel A, but with hnRNP A (UAGGGU/A) and hnRNP H (GGGA) motifs masked. (**C**) Graph of the frequency of non-template strand G-richness as in Panel B, with CpG motifs, hnRNP A (UAGGGU/A) and hnRNP H (GGGA) motifs masked.

DISCUSSION

We have identified a correlation between G-richness near the TSS and pausing in human genes. This correlation emerged from a genome-wide analysis, which examined genes classified as paused in the NCI-60 panel of cell lines or in primary resting T cells. The analysis defined pausing by three different operational criteria: relative levels of transcripts from exon 1 and downstream exons; association of Pol II with the TSS in the absence of transcription; and bivalent histone marks. Downstream but not upstream of the TSS, G-richness of paused genes was biased to the non-template DNA strand. G-rich consensus recognition motifs for sequence-specific DNA or RNA binding proteins, or of CpG dinucleotides, accounted for some but not all G-richness of paused genes. We emphasize that while the correlation between G-richness and pausing was strong, it did not apply to all genes. Additional mechanisms undoubtedly contribute to pausing and G-richness is likely to be only one of the many factors that modulate pausing at any given gene.

The correlation between pausing and G-richness was particularly apparent at the 5′-end of the first intron, where paused genes proved significantly more likely to carry GrIn1 elements, defined as at least one G4 motif within the first 100 bp of the first intron, on the non-template DNA strand (4). GrIn1 elements characterized 57% of paused genes and only 38% of non-paused genes. The genomic position of GrIn1 elements is consistent with a possible role in promoter-proximal pausing. GrIn1 elements lie at the very 5′-end of the first intron, or ~ 200–300 bp downstream of the TSS, as the median distance from the TSS to the 5′-end of the first intron is 198 bp for human genes and GrIn1 elements are about 100 bp in length. Promoter-proximal pausing occurs in the region +20 to +50 relative to the TSS (2). A regulatory element 200–300 bp downstream from the TSS could readily communicate with Pol II or other components of the transcription apparatus, to cause Pol II to pause.

G4 motifs and the mechanism of pausing

The correlation between G4 motifs and pausing suggests that dynamic structures formed upon transcription of a region bearing G4 motifs may contribute to regulation of pausing *in cis*. Figure 6 illustrates those structures, which may promote pausing by distinctive mechanisms: (i) A G4 DNA structure formed behind the advancing polymerase may be recognized by factors that regulate pausing, which in turn cause polymerase to pause. A compelling precedent for a *cis*-regulatory role for G4 DNA has recently been provided by evidence that G4 DNA formation controls pilin gene antigenic variation in *Neisseria gonorrhea* (27). In addition, the human TOP1 gene has recently been found regulated by pausing in the first intron at conserved G4 DNA elements (8). Alternatively, a G4 structure in the DNA might serve as a roadblock to an advancing polymerase, suggested by *in vitro* analysis of transcription on G-rich templates (28), as well as evidence that G4 motifs can block progression of DNA polymerase or even the translation machinery (29–31). In that case, pausing would not occur during the first round of transcription, but after a 'pioneering' round of transcription that enabled a G4 DNA structure to form. (ii) A G4 RNA structure in the 5′-end of the nascent transcript may communicate a pause to the transcription apparatus. This mechanism of pausing has been extensively documented in prokaryotes, where RNA hairpins interact with the polymerase complex to promote pausing at specific sites (32). In human cells, the Trans-Activating Response (TAR) element of the HIV-1 retrovirus has been shown to form a stem–loop structure recognized by Trans-Activator of Transcription (TAT) and associated factors to promote transcription (33). (iii) A stable co-transcriptional RNA/DNA hybrid may communicate a signal for pausing via the RNA processing apparatus or the transcription apparatus. Single molecule imaging has provided dramatic evidence of how co-transcriptional RNA/DNA hybrids can contribute to 'pile-ups' of Pol I actively transcribing the G-rich rDNA in budding yeast (34).

Polymerase pausing is transient (35) and specific regulatory mechanisms may enable a polymerase to exit the paused state. A polymerase that pauses upon encountering a G4 structure could resume transcription upon elimination of that structure, e.g. by a G4 helicase; or if the polymerase/G4 interaction was interrupted by another

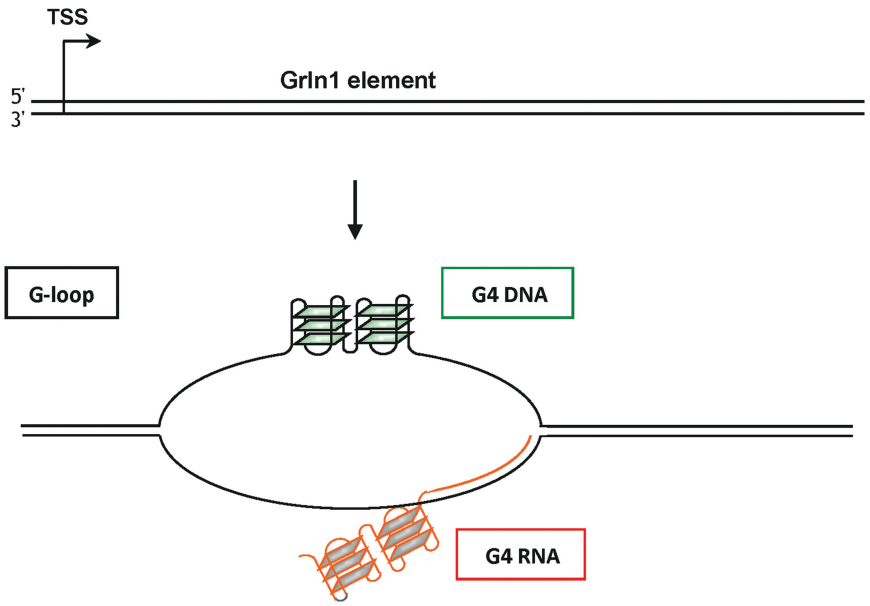

Figure 6. Regulation of transcriptional pausing at G4 motifs. Model of dynamic nucleic acid structures that may contribute to pausing upon transcription of a G-rich region. Mechanisms that contribute to pausing may include: (i) G4 DNA formed behind an advancing polymerase may be recognized by factors that promote pausing, (ii) G4 DNA structure formed in a 'pioneering' round of transcription may serve as a roadblock during the next round of transcription, (iii) a G4 RNA structure in the nascent transcript may communicate a pause to the transcription complex, as occurs in prokaryotes and (iv) a stable co-transcriptional RNA/DNA hybrid may promote pausing, via signals transmitted through the RNA processing apparatus.

factor. In this regard, it is interesting that the hnRNP proteins which interact with RNA in the nucleus contain structural domains (RRM/RBD domains or RGG domains) that recognize and may destabilize G4 structures (36), raising the possibility that they may compete with components of the transcription apparatus for binding to G4 structures.

No single mechanism is likely to account for pausing at every gene. Moreover, the genome-wide analysis that we carried out does not show that all genes that pause carry GrIn1 elements; or that GrIn1 elements are simple identifiers of genes that pause. Nonetheless, the model in Figure 6 should provide a useful starting point for future experiments that elucidate the mechanism of pausing at individual genes and classes of genes.

G-richness and genomic instability in AID-expressing tumors

We have previously shown that G-rich regions are targets of translocations in B cell lymphomas that express the DNA deaminase, AID, although not in T cell leukemias, which do not express AID (11). AID associates with a pausing factor, Spt5 (37). The connection we have established between G-richness and pausing suggests that Spt5 may recruit AID to G-rich paused regions to initiate instability. High levels of AID expression characterize ovarian, breast and prostate malignancies (38) as well as B cell lymphomas. Our results suggest that G-rich sites of pausing may also be targeted for instability in those tumor types.

G4 motifs and drug sensitivity

A role for G4 structures in polymerase pausing has implications for improved understanding of the mechanisms of several classes of drugs, including G4-binding small molecule ligands, G4 aptamers and topoisomerase I poisons. Small molecules that target G4 structures are currently in active development, with telomeres and rDNA as specifically prominent targets (39–42). Our results suggest that interactions with transcription-induced structures may contribute to both the effects and side effects of these drugs. G4 aptamers have also shown promise in treatment of cancer, but their mechanism of action is complex (43). Our results raise the possibility that transcription-induced G4 structures may compete with aptamers for binding key factors, thereby causing unanticipated off-target effects. This could, for example, explain cell type specificity of some aptamers, as binding competition would be determined by the genes expressed in a given cell type.

Camptothecin, a topoisomerase I poison, is the prototype for an important class of cancer chemotherapeutics (44). Treatment of cells with camptothecin has been shown to diminish Pol II pausing (45), an observation which can be explained in terms of the model shown in Figure 6. Formation of co-transcriptional RNA/DNA hybrids is very sensitive to local superhelicity (16,34,46,47). Camptothecin treatment prolongs the half-life of the covalent topoisomerase I/DNA intermediate on the DNA, and may thereby diminish not only local superhelicity but also stability of the local structure containing a co-transcriptional hybrid that promotes pausing. This will contribute to reducing pausing at a subset of genes in camptothecin-treated cells. In this regard, it is interesting that the TOP1 gene, which encodes topoisomerase I, carries a GrIn1 element and is itself regulated by transcriptional pausing (8), which may render TOP1 expression sensitive to local superhelicity, and to camptothecin. The effect of camptothecin on transcript levels is likely to differ from gene to gene, depending on details of local regulation of gene expression and DNA architecture.

ACKNOWLEDGMENTS

We thank members of our laboratories for helpful discussions.

FUNDING

US National Cancer Institute (P01 CA77852 to N.M.); Basic and Cancer Immunology Training Grant (CA009537 to J.E.); US National Institutes of Health (R01 GM41712 and NIH R01 GM65988 to N.M.); Cancer Research Institute Tumor Immunology Predoctoral Training Grant (to J.E.); Intramural Research Program of the National Cancer Institute, Center for Cancer Research support (Z01 BC 006150-19LMP to Y.P. and W.C.R.). Funding for open access charge: P01 NCI CA77852.

Conflict of interest statement. None declared.

REFERENCES

1. Barski,A., Cuddapah,S., Cui,K., Roh,T.Y., Schones,D.E., Wang,Z., Wei,G., Chepelev,I. and Zhao,K. (2007) High-resolution profiling of histone methylations in the human genome. *Cell*, **129**, 823–837.
2. Margaritis,T. and Holstege,F.C. (2008) Poised RNA polymerase II gives pause for thought. *Cell*, **133**, 581–584.
3. Gilmour,D.S. (2009) Promoter proximal pausing on genes in metazoans. *Chromosoma*, **118**, 1–10.
4. Eddy,J. and Maizels,N. (2008) Conserved elements with potential to form polymorphic G-quadruplex structures in the first intron of human genes. *Nucleic Acids Res.*, **36**, 1321–1333.
5. 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.
6. 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.
7. Phan,A.T., Kuryavyi,V. and Patel,D.J. (2006) DNA architecture: from G to Z. *Curr. Opin. Struct. Biol.*, **16**, 288–298.
8. Reinhold,W.C., Mergny,J.L., Liu,H., Ryan,M., Pfister,T.D., Kinders,R., Parchment,R., Doroshow,J., Weinstein,J.N. and Pommier,Y. (2010) Exon array analyses across the NCI-60 reveal potential regulation of TOP1 by transcription pausing at guanosine quartets in the first intron. *Cancer Res.*, **70**, 2191–2203.
9. 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.
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. 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.
12. Vallur,A.C. and Maizels,N. (2008) Activities of human exonuclease 1 that promote cleavage of transcribed immunoglobulin switch regions. *Proc. Natl Acad. Sci. USA*, **105**, 16508–16512.
13. Aguilera,A. (2005) mRNA processing and genomic instability. *Nat. Struct. Mol. Biol.*, **12**, 737–738.
14. Li,X. and Manley,J.L. (2006) Cotranscriptional processes and their influence on genome stability. *Genes Dev.*, **20**, 1838–1847.
15. Lin,Y., Dent,S.Y., Wilson,J.H., Wells,R.D. and Napierala,M. (2010) R loops stimulate genetic instability of CTG.CAG repeats. *Proc. Natl Acad. Sci. USA*, **107**, 692–697.
16. Sordet,O., Redon,C.E., Guirouilh-Barbat,J., Smith,S., Solier,S., Douarre,C., Conti,C., Nakamura,A.J., Das,B.B., Nicolas,E. *et al.* (2009) Ataxia telangiectasia mutated activation by transcription- and topoisomerase I-induced DNA double-strand breaks. *EMBO Rep.*, **10**, 887–893.
17. Hubbard,T., Andrews,D., Caccamo,M., Cameron,G., Chen,Y., Clamp,M., Clarke,L., Coates,G., Cox,T., Cunningham,F. *et al.* (2005) Ensembl 2005. *Nucleic Acids Res.*, **33**, D447–453.
18. 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.
19. Irizarry,R.A., Bolstad,B.M., Collin,F., Cope,L.M., Hobbs,B. and Speed,T.P. (2003) Summaries of affymetrix geneChip probe level data. *Nucleic Acids Res.*, **31**, e15.
20. Ryan,M.C., Zeeberg,B.R., Caplen,N.J., Cleland,J.A., Kahn,A.B., Liu,H. and Weinstein,J.N. (2008) SpliceCenter: a suite of web-based bioinformatic applications for evaluating the impact of alternative splicing on RT-PCR, RNAi, microarray, and peptide-based studies. *BMC Bioinformatics*, **9**, 313.
21. Scherf,U., Ross,D.T., Waltham,M., Smith,L.H., Lee,J.K., Tanabe,L., Kohn,K.W., Reinhold,W.C., Myers,T.G., Andrews,D.T. *et al.* (2000) A gene expression database for the molecular pharmacology of cancer. *Nat. Genet.*, **24**, 236–244.
22. Weinstein,J.N. and Pommier,Y. (2003) Transcriptomic analysis of the NCI-60 cancer cell lines. *C. R. Biol.*, **326**, 909–920.
23. Shoemaker,R.H. (2006) The NCI60 human tumour cell line anticancer drug screen. *Nat. Rev. Cancer*, **6**, 813–823.
24. Bernstein,B.E., Mikkelsen,T.S., Xie,X., Kamal,M., Huebert,D.J., Cuff,J., Fry,B., Meissner,A., Wernig,M., Plath,K. *et al.* (2006) A bivalent chromatin structure marks key developmental genes in embryonic stem cells. *Cell*, **125**, 315–326.
25. Burd,C.G. and Dreyfuss,G. (1994b) 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. Cahoon,L.A. and Seifert,H.S. (2009) An alternative DNA structure is necessary for pilin antigenic variation in Neisseria gonorrhoeae. *Science*, **325**, 764–767.
28. Tornaletti,S., Park-Snyder,S. and Hanawalt,P.C. (2008) G4-forming sequences in the non-transcribed DNA strand pose blocks to T7 RNA polymerase and mammalian RNA polymerase II. *J. Biol. Chem.*, **283**, 12756–12762.
29. Woodford,K.J., Howell,R.M. and Usdin,K. (1994) A novel K(+)-dependent DNA synthesis arrest site in a commonly occurring sequence motif in eukaryotes. *J. Biol. Chem.*, **269**, 27029–27035.
30. 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.
31. Arora,A., Dutkiewicz,M., Scaria,V., Hariharan,M., Maiti,S. and Kurreck,J. (2008) Inhibition of translation in living eukaryotic cells by an RNA G-quadruplex motif. *RNA*, **14**, 1290–1296.
32. Toulokhonov,I., Zhang,J., Palangat,M. and Landick,R. (2007) A central role of the RNA polymerase trigger loop in active-site rearrangement during transcriptional pausing. *Mol. Cell*, **27**, 406–419.
33. Stevens,M., De Clercq,E. and Balzarini,J. (2006) The regulation of HIV-1 transcription: molecular targets for chemotherapeutic intervention. *Med. Res. Rev.*, **26**, 595–625.
34. El Hage,A., French,S.L., Beyer,A.L. and Tollervey,D. (2010) Loss of Topoisomerase I leads to R-loop-mediated transcriptional blocks during ribosomal RNA synthesis. *Genes Dev.*, **24**, 1546–1558.
35. Maizels,N.M. (1973) The nucleotide sequence of the lactose messenger ribonucleic acid transcribed from the UV5 promoter mutant of Escherichia coli. *Proc. Natl Acad. Sci. USA*, **70**, 3585–3589.
36. Maizels,N. (2006) Dynamic roles for G4 DNA in the biology of eukaryotic cells. *Nat. Struct. Mol. Biol.*, **13**, 1055–1059.
37. Pavri,R., Gazumyan,A., Jankovic,M., Di Virgilio,M., Klein,I., Ansarah-Sobrinho,C., Resch,W., Yamane,A., San-Martin,B.R., Barreto,V. *et al.* (2010) Activation-induced cytidine deaminase targets DNA at sites of RNA polymerase II stalling by interaction with Spt5. *Cell*, **143**, 122–133.
38. Pauklin,S., Sernandez,I.V., Bachmann,G., Ramiro,A.R. and Petersen-Mahrt,S.K. (2009) Estrogen directly activates AID transcription and function. *J. Exp. Med.*, **206**, 99–111.
39. Brassart,B., Gomez,D., De Cian,A., Paterski,R., Montagnac,A., Qui,K.H., Temime-Smaali,N., Trentesaux,C., Mergny,J.L., Gueritte,F. *et al.* (2007) A new steroid derivative stabilizes g-quadruplexes and induces telomere uncapping in human tumor cells. *Mol. Pharmacol.*, **72**, 631–640.
40. De Cian,A. and Mergny,J.L. (2007) Quadruplex ligands may act as molecular chaperones for tetramolecular quadruplex formation. *Nucleic Acids Res.*, **35**, 2483–2493.
41. Drygin,D., Siddiqui-Jain,A., O'Brien,S., Schwaebe,M., Lin,A., Bliesath,J., Ho,C.B., Proffitt,C., Trent,K., Whitten,J.P. *et al.* (2009) Anticancer activity of CX-3543: a direct inhibitor of rRNA biogenesis. *Cancer Res.*, **69**, 7653–7661.
42. Sparapani,S., Haider,S.M., Doria,F., Gunaratnam,M. and Neidle,S. (2010) Rational design of acridine-based ligands with selectivity for human telomeric quadruplexes. *J. Am. Chem. Soc.*, **132**, 12263–12272.
43. Reyes-Reyes,E.M., Teng,Y. and Bates,P.J. (2010) A new paradigm for aptamer therapeutic AS1411 action: uptake by macropinocytosis and its stimulation by a nucleolin-dependent mechanism. *Cancer Res.*, **70**, 8617–8629.
44. Pommier,Y., Leo,E., Zhang,H. and Marchand,C. (2010) DNA topoisomerases and their poisoning by anticancer and antibacterial drugs. *Chem. Biol.*, **17**, 421–433.
45. Khobta,A., Ferri,F., Lotito,L., Montecucco,A., Rossi,R. and Capranico,G. (2006) Early effects of topoisomerase I inhibition on RNA polymerase II along transcribed genes in human cells. *J. Mol. Biol.*, **357**, 127–138.
46. Hraiky,C., Raymond,M.A. and Drolet,M. (2000) RNase H overproduction corrects a defect at the level of transcription elongation during rRNA synthesis in the absence of DNA topoisomerase I in Escherichia coli. *J. Biol. Chem.*, **275**, 11257–11263.
47. Drolet,M., Broccoli,S., Rallu,F., Hraiky,C., Fortin,C., Masse,E. and Baaklini,I. (2003) The problem of hypernegative supercoiling and R-loop formation in transcription. *Front. Biosci.*, **8**, d210–d221.