Nucleic Acids Research, 2023, **51**, 12124–12139
https://doi.org/10.1093/nar/gkad948
Advance access publication date: 1 November 2023
Gene regulation, Chromatin and Epigenetics

Genetic variations in G-quadruplex forming sequences affect the transcription of human disease-related genes

Agustín Lorenzatti[1,†], Ernesto J. Piga[1,†], Mauro Gismondi[2], Andrés Binolfi[1,3], Ezequiel Margarit[2], Nora B. Calcaterra[1] and Pablo Armas[1,*]

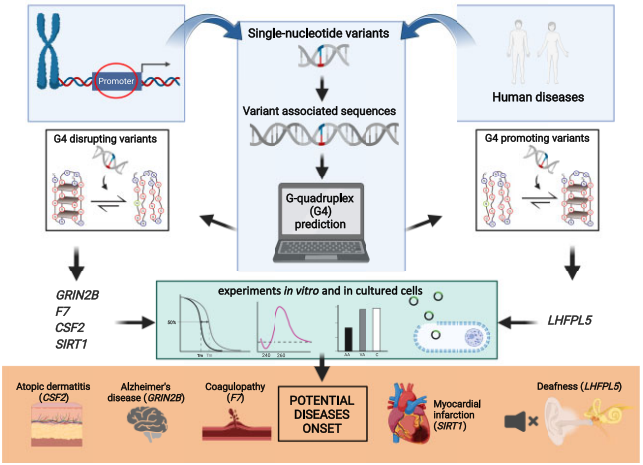
[1]Instituto de Biología Molecular y Celular de Rosario (IBR), Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET) - Facultad de Ciencias Bioquímicas y Farmacéuticas, Universidad Nacional de Rosario (UNR), Ocampo y Esmeralda, Rosario S2000EZP, Santa Fe, Argentina
[2]Centro de Estudios Fotosintéticos y Bioquímicos (CEFOBI), Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET) - Facultad de Ciencias Bioquímicas y Farmacéuticas, Universidad Nacional de Rosario (UNR), Suipacha 531, Rosario, Santa Fe, Argentina
[3]Plataforma Argentina de Biología Estructural y Metabolómica (PLABEM), Ocampo y Esmeralda, Rosario S200EZP, Santa Fe, Argentina

[*]To whom correspondence should be addressed. Tel: +54 341 4237070 (Ext 654); Fax: +54 341 4237070 (Ext 607); Email: armas@ibr-conicet.gov.ar
[†]The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

Abstract

Guanine-rich DNA strands can fold into non-canonical four-stranded secondary structures named G-quadruplexes (G4s). G4s folded in proximal promoter regions (PPR) are associated either with positive or negative transcriptional regulation. Given that single nucleotide variants (SNVs) affecting G4 folding (G4-Vars) may alter gene transcription, and that SNVs are associated with the human diseases' onset, we undertook a novel comprehensive study of the G4-Vars genome-wide (G4-variome) to find disease-associated G4-Vars located into PPRs. We developed a bioinformatics strategy to find disease-related SNVs located into PPRs simultaneously overlapping with putative G4-forming sequences (PQSs). We studied five G4-Vars disturbing *in vitro* the folding and stability of the G4s located into PPRs, which had been formerly associated with sporadic Alzheimer's disease (*GRIN2B*), a severe familiar coagulopathy (*F7*), atopic dermatitis (*CSF2*), myocardial infarction (*SIRT1*) and deafness (*LHFPL5*). Results obtained in cultured cells for these five G4-Vars suggest that the changes in the G4s affect the transcription, potentially contributing to the development of the mentioned diseases. Collectively, data reinforce the general idea that G4-Vars may impact on the different susceptibilities to human genetic diseases' onset, and could be novel targets for diagnosis and drug design in precision medicine.

Graphical abstract



Introduction

Under certain conditions, specific DNA sequences do not adopt the canonical B- structure, but instead form non-B (non-canonical) DNA secondary structures such as cruciform, hairpin, i-motif, Z-DNA and G-quadruplex (G4) (1), among others. G4s are dynamic structures formed by the self-folding of G-rich single-stranded DNA exposed during replication and transcription as a consequence of the negative supercoiling associated with these processes (2). Putative G4-forming sequences (PQSs) have been identified within the genomes of

Received: August 12, 2022. **Revised:** September 22, 2023. **Editorial Decision:** September 29, 2023. **Accepted:** October 12, 2023

all kingdoms of life (3–5) and even in viral genomes (6), thus suggesting an important role of these structures throughout evolution.

The core structure of a G4 consists of a planar arrangement of four guanine residues bound by Hoogsteen base pairing conforming G-quartets or G-tetrads. The stacking of at least two G-tetrads establishes the G4, which is in turn stabilised by the coordination of monovalent cations, mainly by K^+. A widely accepted consensus PQS is $G_{\geq 3}$ N_{1-7} $G_{\geq 3}$ N_{1-7} $G_{\geq 3}$ N_{1-7} $G_{\geq 3}$, wherein G-tracks are connected by loops of varying length and nucleotide (N) composition (7). However, non-canonical G4s can also form within sequences that have longer loops or less than three guanines per repeat (5,8). The numbers of guanines per G-track and the length of the loops influence the stability of the G4 structure (9).

G4s were initially described in telomeric DNA sequences (10); but later, a significant number of studies have shown that G4s are important regulators of multiple cellular processes, such as replication, transcription and genome maintenance (2,11). Although the relevance of G4 structures in living cells was questioned in the past, accumulating experimental data now support the existence and importance of these structures in living cells and organisms (5,12). The role of G4s in cancer is an area of intense study (13). G4s are found in telomeric DNA (14–16) and telomeric non-coding RNA (17), and play key roles in replication and genome instability (18,19). Besides, most promoters of oncogenes harbour more G4 motifs than promoters of regulatory or tumour suppressor genes (20). Apart from the exhaustive studies that are being carried out to elucidate the role of G4s in the development of cancer, the exploration of the role of G4s in the context of other human diseases has grown-up remarkably in the last years (21). G4s have been initially described to function as inhibitors of transcription, acting as roadblocks impeding RNA polymerase mediated transcription (2). However, other more recent studies have added substantial evidence showing that G4s are globally correlated with transcriptional enhancement rather than repression (although there may be individual cases where G4s are repressive), and showed that the role of G4 structures in transcription is more complex than a simple 'on/off' switch (22). For understanding the unpredictable and variable effect of G4s on transcriptional control, it is important to consider the G4s not as isolated entities within a specific genomic location, but as structures coexisting with other biomolecules in living cells as part of an interconnected network. Novel proposed mechanisms for G4-mediated transcriptional regulation involve G4s functioning as binding hubs for many transcription factors; as mediators of histone marks and interacting with chromatin remodelling proteins, thus shaping the chromatin architecture; as influencers of the formation and stability of transcriptional loops including R-loops and long-range enhancer-promoter loops; among others (22).

Advances in the human genome sequencing during the last decades have shown sequence variations among individuals. Most of these variations consist of short indels and single-nucleotide variants (SNVs), which occur on average every 4–5 bp (https://www.ncbi.nlm.nih.gov/snp/). Genome-wide association studies (GWAS) have identified >200 000 SNVs loci associated with several human diseases or traits. The majority of these SNVs are located in non-coding regions, including intergenic and intronic regions, and could affect gene expression by overlapping with transcription or translation regulatory elements (23). Therefore, one main challenge in human genetics is to understand the biological mechanisms by which SNVs influence phenotypes, including disease risk (24). The existence of 5 million gains/losses or structural conversions of G4s generated by the presence of SNVs has been recently reported. Of these, 3.4 million are within genes, preferentially enriched near the transcription start site (TSS) and overlapping with transcription factor binding sites and enhancers (25). Bearing in mind that G4s and the transcription factors associated with them cooperate to specify differential transcriptional programs (26), gains/losses or structural conversions of G4s generated by the presence of SNVs could not only alter a G4 structure but also influence gene expression among individuals (25,27,28). Given that the formation of secondary structures, such as G4s, may play a pivotal role in the occurrence of SNVs (29–31), it is plausible to view G4s as potential contributors of genetic diversity. This suggests a synergistic and/or cooperative interplay between G4 formation and the subsequent diversification of G4 structural motifs (32). In coincidence, an integrative analyses of somatic mutations (copy-number aberrations and SNVs) found in patient-derived tumour cells, alongside with differentially G4-enriched genomic regions, has associated G4s with cell-type specific transcriptional control, intra-tumour heterogeneity and cancer cells genome mutagenesis and instability (32,33).

A couple of works have reported a connection between disease risk and the gain/loss of G4s caused by SNVs. One of them reports that an SNV stabilising a G4 in the first intron of the calcium channel gamma subunit 8 gene (*CACNG8*) suppresses gene expression thus increasing the susceptibility to antisocial personality disorder (ASPD) (34). In the other, it was shown that the SNV responsible for the decrease in the transcription of the gene encoding placental anticoagulant protein annexin A5 (*ANXA5*) diminishes the potential for G4 formation *in vitro* and *in vivo*, suggesting an association between loss of the G4 and obstetric complications (35). Thus, we wondered whether gains/losses or structural conversions of G4s caused by disease-associated SNVs located in the proximal promoter regions (PPRs) of human genes influence their transcriptional expression. If so, not only the presence of a G4 but also its structural variants could impact on disease onset or susceptibility.

Here we report a novel comprehensive bioinformatics analysis performed by searching disease-related SNVs within PQSs located into the PPRs of human genes. We found novel PQSs that were characterised *in vitro* as being able to fold as G4s. In addition, we identified five SNVs associated with diseases that either promote or impair *in vitro* the formation and stability of G4s located into the PPRs of *GRIN2B*, *SIRT1*, *CSF2*, *F7* and *LHFPL5* genes. Importantly, we gathered experimental data showing that SNVs present into the PPRs of these five genes affect the transcription of a reporter gene in cultured cells. This is especially relevant for the SNVs of *GRIN2B* and *F7*, which were previously suggested to be responsible for altering the transcription of these genes and being associated with the related diseases (36,37). Furthermore, the action of specific G4 ligands on the transcription of the endogenous *F7* gene in cultured cells is in agreement with the results shown here in our experiments with reporter genes and with data previously reported by other authors (37). Taken together, the data reinforce the general idea that SNVs within G4s contribute to the onset or susceptibility of human pathologies.

Materials and methods

Obtaining human genomic variants and flanking sequences

Human Short Variants (HSV, including SNVs and indels excluding flagged variants) corresponding to HGMD-PUBLIC, ClinVar and dbSNP databases, and Human Somatic Short Variants (HSSV, including SNVs and indels excluding flagged variants) corresponding to dbSNP, ClinVar and COSMIC databases were downloaded from Ensembl Variation database (human reference genome GRCh38.p12; releases 92 and 93; https://www.ensembl.org/index.html) (38) using the Biomart tool. HSV and HSSV datasets were filtered to obtain disease-associated variants located into PPRs, defined here as the region spanning 1000 bp upstream from the TSSs reported in Ensembl. The 50 nucleotides up and downstream sequences flanking the variant positions were downloaded as multi-fasta files. In this work, and according to Ensembl, the 'ancestral allele' (AA) was defined as the allele found in closely related species and is thought to reflect the allele present at the time of speciation. On the other hand, 'variation allele' (VA) refers, in this work, to variation sequences derived from the ancestral ones. The nucleobases of both kinds of alleles (the AA and VA) were included as headers for every downloaded variant-flanking sequence. Separated multi-fasta sequence files were generated with a custom Perl script (Supplementary File S1), thus allowing the specific substitution of the variant position with the nucleobase corresponding to either the AA or VA, and the generation of two new multi-fasta files containing the substituted sequences; one of these new files represents AA-sequences and the other one the respective VA-sequences (Figure 1A). Since only the AA was reported for COSMIC and HGMD-PUBLIC variants, for the AA-sequences obtained from these databases we generated the other three possible VA by replacing the AA-informed nucleobase with the other three possible nucleobases.

PQSs finding, variant id intersections and assessment of G4 formation propensity

A Perl script (Supplementary File S2) was generated to search for PQSs in both strands of the downloaded DNA sequences. We used the extended PQS definition consisting of four tracks of 3–7 guanines (G) (or cytosines (C) to consider the complementary strand) separated by three loops of 1–12 nucleotides (A, C, G or T), according to Sahakyan et al. (39). The script reports the results in a tabular output file with the variant identification numbers (ids) contained in PQSs, the found PQSs and the number of PQSs found per sequence. PQSs searches were performed on multi-fasta files generated for AA- and VA-sequences.

Variants occurring within PQSs (hereafter named as pG4-Vars) were analysed to identify those ones promoting or disrupting PQSs occurrence. Variant ids list intersections were performed by using Venny software (https://bioinfogp.cnb.csic.es/tools/venny/). Quadron package (http://quadron.atgcdynamics.org/) (39) was implemented on multi-fasta files containing the variant-flanking sequences (150 nucleotides both upstream and downstream, as required by the software) of the pG4-Vars downloaded from Ensembl. pG4-Vars within PQSs with Quadron Q parameter ≥19 were named here as G4-Vars. Variant ids corresponding to identified G4-Vars after Quadron predictions were used as inputs of the Biomart tool to download associated phenotypes.

Oligonucleotides and compounds

The oligonucleotides representing the sequences studied in this work contained the complete PQSs flanked both at 5′ and 3′ ends by five additional nucleotides corresponding to the reference sequences informed in the Ensembl genome browser. For each case, mutations in PQSs were designed based on G-tracks disruption for impending G4 formation and were tested using the QGRS Mapper software (https://bioinformatics.ramapo.edu/QGRS/index.php) (40). Synthetic single-stranded oligodeoxyribonucleotides (Supplementary Table S1) were purchased from Macrogen Inc. with cartridge purification, dissolved in bi-distilled water and stored at −20°C until use. Concentrations were determined by spectrophotometry using extinction coefficients provided by the manufacturer.

Circular dichroism (CD) spectroscopy

Intramolecular G4s were folded by dissolving 2 μM oligodeoxyribonucleotides in 10 mM Tris–HCl pH 7.5 and different KCl concentrations, as indicated in each figure, heating for 5 min at 95°C and slowly cooling to 20°C for at least 2 h. CD spectra were recorded at 20°C covering a wavelength range of 220–320 nm with a Jasco J-1500 spectropolarimeter (10 mm quartz cell, 1 nm band width, 1 nm data pitch, 100 nm/min scanning speed, 1 s response time). The recorded spectra represent a smoothened average of four scans, zero-corrected at 320 nm. The absorbance of the buffer was subtracted from the recorded spectra.

BG4 dot-blots

Dots were spotted using 10 μl of 1 μM folded oligodeoxiribonucleotides (in 10 mM Tris pH 7.5 and 100 mM KCl) on Hybond™-N + membrane (Amersham) and cross-linked with a UV lamp (UVP, 302 nm) during 5 min. Membranes were washed twice with Tris-Buffer Saline with 0.1% Tween 20 detergent (TBS-T) for 10 min at room temperature and then blocked in TBS-T containing 5% blotto for 1 h. Membranes were first incubated overnight at 4°C with 1 μg/ml of recombinant BG4-antibody (41). Then, membranes were incubated with Anti-Flag antibody (clone M2, Sigma Aldrich Catalogue Number F3165; 1 μg/ml) for 1 h at room temperature. Finally, membranes were incubated for 1 h at room temperature with anti-mouse conjugated to Horseradish Peroxidase (Jackson ImmunoResearch Inc., Code: 115–035-003, dilution 1/25 000). After each antibody incubation and before development, membranes were washed with TBS-T five times during 10 min at room temperature. Finally, G4s detection was developed by chemiluminescence (Biolumina Chemiluminescent Substrate, Kallium Technologies, CABA, Argentina) followed by exposition on X-ray films (Amersham Hyperfilm ECL, GE, MA, USA).

1D ^{1}H Nuclear magnetic resonance (NMR)

Spectra were acquired at 20°C on a 700 MHz Bruker Avance III spectrometer equipped with a triple resonance inverse NMR probe (5 mm 1H/D-13C/15N TXI). Samples containing 50 μM oligodeoxyribonucleotides folded as G4s (as described for CD spectroscopy) in the presence of KCl, as indicated in each figure, were loaded into 5 mm Shigemi tubes. 1D ^{1}H NMR spectra were registered using a pulse sequence with excitation sculpting (zgesgp) for water suppression (42). 8K points, 1K scans, a recycling delay of 1.4 s and a sweep

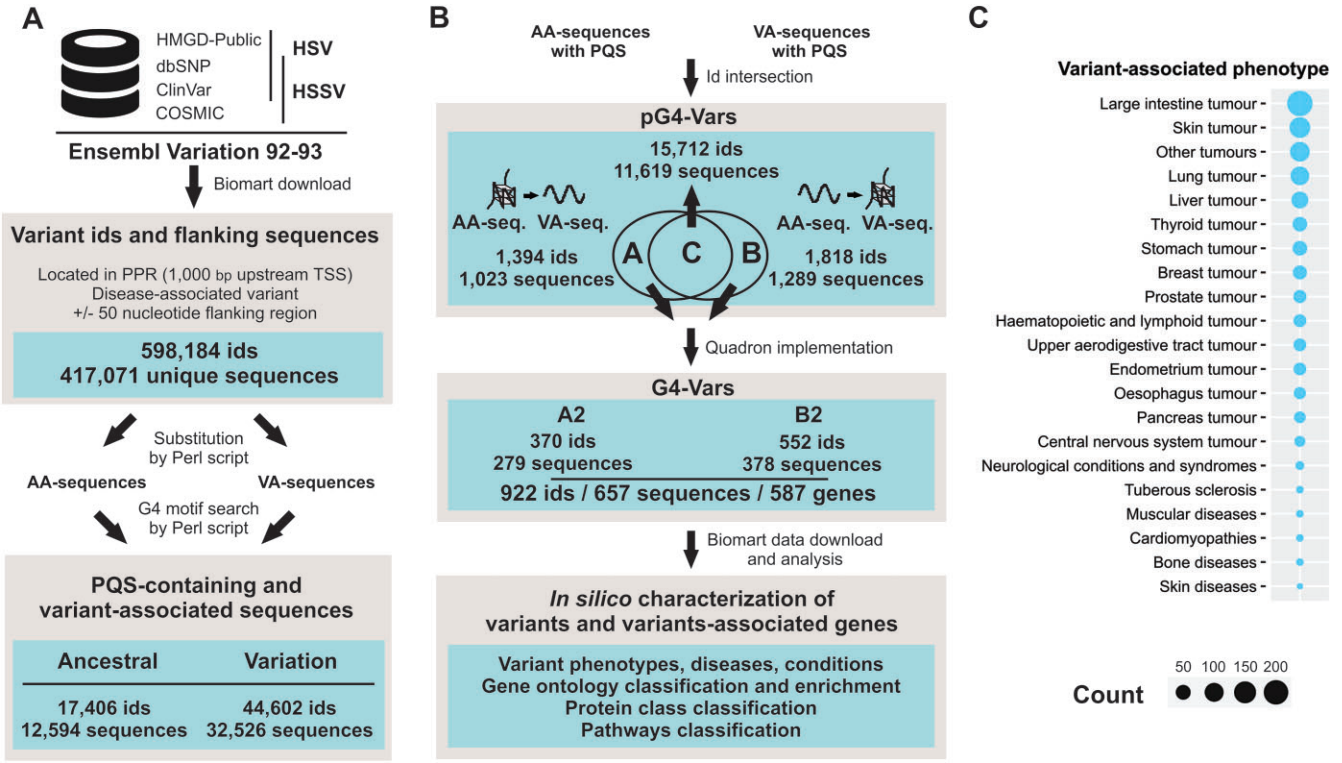


Figure 1. Bioinformatics pipeline towards the identification of disease-related G4-Vars. **(A)** Variant-associated sequences were downloaded from the Ensembl Variation database. A custom Perl script was designed to generate AA- and VA-sequences. Another custom Perl script was used to search for PQSs on both strands of AA- or VA-sequences. **(B)** Intersections between id lists of AA- and VA-sequences containing PQSs allowed identifying groups A and B of pG4-Vars. Quadron implementation yielded sub-groups A2 and B2 of G4-Vars (from groups A and B, respectively). Finally, variant phenotypes and variant-associated genes were characterised according to Gene Ontology (GO) terms, Panther database annotations and associated human diseases or conditions. Numbers of ids, sequences and genes obtained in each step of the pipeline (grey boxes) are shown (light blue boxes). **(C)** Computational characterization of G4-Vars associated phenotypes. Circles sizes represent the number of variants ids, scale is shown below the figure.

width of 18 ppm were used. Experimental time for each NMR spectrum was 30 min. Processing was done using an exponential window function multiplication with a line broadening of 10 Hz and baseline correction. Topspin 3.7 software (Bruker, Biospin) was used for acquisitions, processing and analysis of the NMR spectra.

qPCR stop assays (qPSA)

qPSA were performed as reported elsewhere (43), with some modifications. When synthetic oligodeoxyribonucleotides were used as templates, each reaction tube (10 µl) consisted of 0.5× SYBR Green, 250 nM of each PCR primer (Supplementary Table S1), 5 pM of template oligodeoxyribonucleotides (AA, VA or M, Supplementary Table S1), 0.5 units of Platinum *Taq* DNA polymerase (Invitrogen), the PCR reaction buffer provided by the manufacturer, 2.5 mM of $MgCl_2$ and 200 µM of each dNTP. Reactions were performed in a Realplex 4 Thermocycler (Eppendorf) using the following program: 95°C for 3 min (1 cycle) followed by a 2-step reaction of 95°C for 5 s and 60°C for 20 s (40 cycles), and finally a 20 min melting curve. When plasmids where used as templates, 500 ng of the pGL3-promoter vector (Promega), or the constructions derived from it, were dissolved in 10 mM Tris–HCl pH 7.5 and 100 mM KCl, heated for 5 min at 95°C and slowly cooled to 20°C for at least 2 h, and then used as templates. Each reaction tube (15 µl) consisted of 0.75× HOT FIREPol EvaGreen qPCR Mix Plus (ROX) and 250 nM of each PCR primer (Supplementary Table S1). Two reactions were performed for each plasmid, one using a pair of primers annealing to the flanking sequences of the SmaI restriction site of pGL3-promoter vector plasmid, and another using a pair of primers annealing to the coding region of firefly luciferase; this latter reaction was used as an internal reference for quantification. For pre-transfection assays, 1 µl of 0.2 ng/µl plasmids were used as templates. For post-transfection assays, 1 µl of 1/100 dilution of cell lysates were used as templates. Reactions were performed in a Realplex 4 Thermocycler (Eppendorf) using the following program: 95°C for 12 min (1 cycle) followed by a 2-step reaction of 95°C for 5 s and 60°C for 20 s (40 cycles) and finally a 20 min melting curve. Three technical replicates were performed for each experimental condition. Three independent qPSAs were done for each case. The validity of the qPCR data was assured by following the MIQE guidelines (44).To express the relative DNA amplification $2^{-\Delta Cq}$ values were determined, where ΔCq was defined as threshold cycle for each reaction minus threshold cycle for the reference reaction.

Plasmids constructions

Duplex DNAs were generated by annealing oligodeoxiribonucleotides containing AA- or VA-sequences with the corresponding complementary strands (Supplementary Table S1) and then cloned using SmaI restriction site upstream the basal SV40 promoter into pGL3-promoter vector plasmid (Promega), in the same strand (coding or template) as found in the genome, as previously described (12). Plasmids were

sequenced and used for cell culture transfections as described below.

Cell culture and transfection

Human embryonic kidney 293 (HEK-293) cells were grown in Dulbecco's modified Eagle's medium (DMEM) and supplied with 10% foetal bovine serum (FBS), in a humidified cell incubator with an atmosphere of 5% CO_2 at 37°C. Cells were seeded at a density of 3×10^6 cells per 35 mm dish 24 h prior to transfection in DMEM containing 10% FBS. Cells were transfected using calcium phosphate method (45). For luciferase reporter assays (LRAs), 250 ng of the pGL3-promoter vector (or the constructions derived from it) and 125 ng of the pCMV-β-gal plasmid (Promega) were used for transfection. The transfection medium was replaced by fresh medium 5 h after transfection, and cells were grown as indicated above for 48 h post-transfection (hpt). For qPSAs, 500 ng of the pGL3-promoter vector (or the constructions derived from it) were dissolved in 10 mM Tris–HCl pH 7.5 and 100 mM KCl, heated for 5 min at 95°C and slowly cooled to 20°C for at least 2 h, and then used for transfection. The transfection medium was replaced by fresh medium 5 h after transfection, and cells were grown as indicated above and at 24 hpt a crosslinking was performed by adding 100 µl of formaldehyde 16% w/v into the growing medium for 10 min, and reaction was stopped by adding glycine to 0.12 M final concentration for 10 min. Cells were washed with phosphate buffer saline (PBS) and lysated using 250 µl of lysis buffer from Luciferase Reporter Assay System (Promega, USA).

Luciferase reporter assays (LRAs)

LRAs were performed as previously described (12) with few modifications. At 48 hpt, cells were collected, lysated and firefly luciferase (FL) activities were measured using Luciferase Reporter Assay System (Promega, USA). FL activities were normalised to β-gal activities and expressed as the ratio of FL/βgal. Values determined for each construct were relativized to those for AA constructs for each PQS in study. The normalised ratio was obtained from three independent biological replicates for each experiment, and experiments were repeated three times.

Results

Obtaining a novel dataset of disease-related G4-Vars within PPRs

Short genetic variants (SNVs) may cause the gain, loss, and/or change in folding propensity of G4s that modulate transcriptional gene expression (25). Therefore, it is a fundamental issue to address the physiological function and pathological implications that such SNVs may represent. In view of this, we developed a bioinformatic pipeline (Figure 1) with the purpose of identifying disease-related SNVs either promoting or disrupting the formation of G4s. We downloaded the disease-related HSV and HSSV contained into PPRs (defined as 1000 bp upstream the TSS), along with their 50 bp-flanking sequences reported in the databases included in Ensembl Variation database (Figure 1A and Supplementary Table S2). Among the 598 184 downloaded sequences, 417 071 were unique sequences, as some of them were repeated due to being represented by different variant ids. From the 417 071 sequences, we generated two new sets of sequences, one corresponding to the ancestral allele (AA)-sequences and another to the variation allele (VA)-sequences. Both sets of sequences were used as inputs for a Perl script designed to search for an extended PQS definition consisting of four G-tracks of at least 3 Gs interspersed with extended loops of 1 to 12 nucleotides ($G_{\geq 3}$ $N_{1–12}$ $G_{\geq 3}$ $N_{1–12}$ $G_{\geq 3}$ $N_{1–12}$ $G_{\geq 3}$), which accounts for 2/3 of the experimentally observed G4s (39) (Figure 1A and Supplementary Tables S3 and S4).

We identified 17 106 ids that resulted in AA-sequences containing PQSs corresponding to 12 594 unique sequences. On the other hand, we identified 44 602 ids that resulted in VA-sequences containing PQSs corresponding to 32 526 unique sequences (Figure 1A). For a better understanding, hereafter SNVs occurring within PQSs are named as pG4-Vars. The intersection between both id lists resulted in three groups (Supplementary Table S5). Group A (Supplementary Table S6 and Supplementary File S3) contains ids with PQSs only in the AA-sequences, group B (Supplementary Table S7 and Supplementary File S4) contains ids with PQSs only in the VA-sequences (Figure 1B), and group C, which contains the majority of the ids, includes ids presenting PQSs both in the AA- and VA-sequences. In the case of group C (Supplementary Table S5), the different alleles did not generate PQSs gains/losses. One possible explanation is that SNVs modify the sequence of the loops, the flanking regions or even the G-tracks, but do not affect the PQS consensus. Therefore, sequences belonging to group C were not further assessed in this study.

The 2 312 unique sequences contained in groups A and B (1 023 in group A + 1 289 in group B) were further characterised in order to classify them according to the propensity of G4 formation *in vitro*. For this purpose, we used Quadron algorithm (39), which predicts DNA G4 formation and stability using a machine-learning approach based on experimental data from human genome G4-seq. pG4-Vars within PQSs displaying Quadron scores higher than 19 (hereafter named as G4-Vars) were grouped in two new sub-groups, A2 and B2 (Figure 1B). Sub-group A2 comprises sequences containing PQSs displaying high propensity to fold as G4 only in the AA-sequences, while sub-group B2 consists of sequences containing PQSs with high propensity to fold as G4 only in the VA-sequences. Overall, 922 ids (72 from the HSV and 850 from the HSSV datasets) corresponding to 657 unique sequences were identified as disease-related G4-Vars located into PPRs (Figure 1B and Supplementary Table S8). The G4-Vars identified in sub-groups A2 and B2 represent a sub-set of G4-Vars with potential roles in disease-associated gene transcriptional regulation. However, as we did not select the G4-Vars in function to their link with transcriptional changes, the effects of G4-Vars need to be investigated for each case and experimentally validated by assaying their biochemical effect on G4s formation and their role on transcriptional activity in a cellular context to enable an understanding of the significance of selected variations.

Noteworthy, only four ids from the 370 found in sub-group A2 and only one from the 552 found in sub-group B2 are classified by Ensembl as variants located into PPRs (described as 'upstream gene variant' by Ensembl) (Supplementary Figure S1). Instead, most variants are classified by Ensembl with calculated variant consequences overlapping with transcripts (5′ UTR, coding sequences, 3′ UTR and introns). The low number of G4-Vars found located into PPRs may be explained at least by three possible scenarios. First, Ensembl reports all

the possible transcripts for each gene, which are expressed in different tissues or developmental stages, but may not be the most abundant or representative transcripts. So, in our analyses, many G4-Vars located upstream of a TSS of a gene with several alternative TSSs may be classified by Ensembl as located within longer transcripts. Second, some G4-Vars may overlap with the transcription unit of another gene, being the variant consequence within the transcription unit probably prevalent over the upstream gene variant consequence. Third, an important number of G4-Vars are classified as overlapping with non-coding transcripts, thus they could be located within PPRs overlapping with non-coding transcripts, being the non-coding transcripts variant consequence prevalent over the upstream gene variant consequence. Therefore, it would be expected that more than the five G4-Vars described here as 'upstream gene variant' are relevant for transcriptional control. It is worth mentioning that the SNVs located both into the PPR and simultaneously downstream the TSSs of an alternative transcript (within the transcriptional unit) may represent a complication in the analysis of consequences on gene expression due to putative effects of the G4-Vars on both transcriptional and post-transcriptional regulation of different transcripts for the same gene.

Analysing the variant's clinical significance (i.e. the impact on disease described in the human Ensembl Variation database) of the G4-Vars classified in sub-groups A2 and B2, we observed that only 47 from A2 sub-group and 19 from B2 sub-group have associated clinical significance terms, and only 1 from A2 sub-group and 3 from B2 sub-group presented Clinical significance terms corresponding to 'pathogenic' or 'likely pathogenic' (Supplementary Table S9). The low number of G4-Vars identified suggests that either most of them are not associated with diseases or that they have not yet been classified in the human Ensembl Variation database according to their relevance to the related diseases. In this context, an analysis mainly based on clinical significances could overlook G4-Vars with effects on the fine-tuning of gene expression and, consequently, be relevant for disease predisposition in synergy with other risk factors. Hence, the G4-Vars were assessed for their variant-associated phenotypes in the Ensembl Variation database. We found 833 ids (out of 922) related to a variety of pathologies, the most represented being different kinds of tumours (intestinal, skin, lung, liver, stomach and breast tumours, among others); although other conditions or diseases, such as neurological conditions and syndromes, cardiomyopathies, muscular, bone and skin diseases were also found (Figure 1C). In addition, 922 variant ids were associated with 587 unique genes (according to the Ensembl Genes Database), which were characterised by considering gene ontology (GO) terms (Supplementary Figure S2A), Panther protein class (Supplementary Figure S2B) and pathway (Supplementary Figure S2C). Additionally, a GO enrichment study performed over the entire set of 587 genes showed that biological process terms related to development at different levels are prevalent (Supplementary Figure S2D and Supplementary Table S10).

Selection of G4-Vars for experimental analyses

As our main goal was to identify G4-Vars affecting transcriptional control, and most of the identified G4-Vars were classified by Ensembl with variant consequences overlapping with transcripts, we decided to focus our work on the analyses of the only five cases for which G4-Vars had been described as 'upstream gene variant' located into PPRs and upstream the TSSs of the most biologically relevant transcripts reported in Ensembl: *GRIN2B*, *SIRT1*, *CSF2, F7* and *LHFPL5* genes. *GRIN2B* encodes the glutamate ionotropic receptor NMDA type subunit 2B, a subunit of the *N*-methyl-D-aspartate (NMDA) receptor ion channels. Naturally occurring mutations within this gene are associated with neurodevelopmental disorders (46). The VA of the SNV causing the G4-Var identified here (Table 1 and Supplementary Figure S3A) causes an increase in *GRIN2B* transcriptional expression and was associated with reduced risk of sporadic Alzheimer's disease (SAD) (36). *F7* encodes the Factor VII (FVII) protein, the serine-protease that triggers blood coagulation. A reduced plasma level of FVII results in bleeding diathesis of variable severity. The VA of the SNV causing the G4-Var identified here (Table 1 and Supplementary Figure S3B) causes a reduction in *F7* gene transcriptional expression and is responsible for a severe bleeding familiar disorder due to factor VII deficiency (37). *CSF2* encodes the cytokine colony stimulating factor 2, which has been related with pulmonary alveolar proteinosis and mucositis. The SNV causing the G4-Var identified here (Table 1 and Supplementary Figure S3C) is associated with reduced severity in atopic dermatitis (47), although there is no reported association of this variant with changes in transcriptional expression. *SIRT1* encodes SIRTUIN 1, a member of a family of NAD-dependent class III deacetylases that modulate chromatin function and has been connected to many cellular processes such as cell cycle, response to DNA damage, metabolism, apoptosis and autophagy. *SIRT1* is implicated in different human diseases and the SNV causing the G4-Var identified here (Table 1 and Supplementary Figure S3D) is associated with increased risk of myocardial infarction (48), although there is no reported association of this variant with changes in transcriptional expression. *LHFPL5* encodes LHFPL tetraspan subfamily member 5, a member of the lipoma HMGIC fusion partner (LHFP) family, a subset of the superfamily of tetraspan transmembrane proteins. Mutations in this gene result in deafness, and it is proposed to function in the inner ear as a component of the hair cell's mechanotransduction machinery (49). The SNV causing the G4-Var identified here (Table 1 and Supplementary Figure S3F) is associated with deafness (50) (https://www.ncbi.nlm.nih.gov/clinvar/RCV000288173/), although there is no reported association of this variant with changes in transcriptional expression.

GRIN2B, *SIRT1*, *CSF2* and *F7*, G4-Vars were identified as potentially disrupting the formation of G4s (sub-group A2) and are described in Ensembl as 'upstream gene variant'. In the case of *LHFPL5*, it was identified as potentially promoting the formation of G4s (sub-group B2) and, although the G4-Var is classified in Ensembl as '5 prime UTR variant', is also located upstream of the TSS of the most biologically relevant transcript (main functional isoform, most conserved, highly expressed, and that has the longest coding sequence) (Figure 2A, Table 1 and Supplementary Figure S3F). An additional computational analysis was performed on the sequences corresponding to these five genes (i.e. AA-, VA- and mutated (M)-sequences, in which G to A replacements disrupt the PQS motif, Supplementary Table S1) in order to analyse the incidence of the G4-Vars using other predictors of

A

Sub-Group	Gene	Sequence (5′- 3′)	Strand	Quadron score	Distance to TSS	PQS length
A2	*GRIN2B*	GTGCT**GGGG**AAGT**GGGG**TGGTAA**GGGGG**ATGGAA**GGGGG**ACGGAA**GGGGG**AAGAA	Template	29.65	-128/-528	45
	F7	TCCAA**GGG**TGACAGA**GGG**AT**GG**G**GG**A**GGGG**AGGAC	Template	20.17	-40/-43	25
	CSF2	TCAGA**GGGGG**T**G**G**GG**CT**GGG**TCAAGTC**GGG**ACCCA	Template	27.4	-640/-640	25
	SIRT1	CGCGT**GGG**TGGC**GGG**AGCGCCGAGA**GGG**C**GGGG**G**C**GGCGAT**GGGG**C**GGG**TCACGTGAT**GGGG**TTTAA	Coding	23.24	-61/-61	57
B2	*LHFPL5*	CGGAC**GG**G**GG**A**GGGGG**C**GGGG**CTCTC**GGG**AGCCGTGAGCC**GGG**AAGA**GGG**AGAC**GGG**CA**GGG**CGGCG	Coding	28.53	-16/-16	57

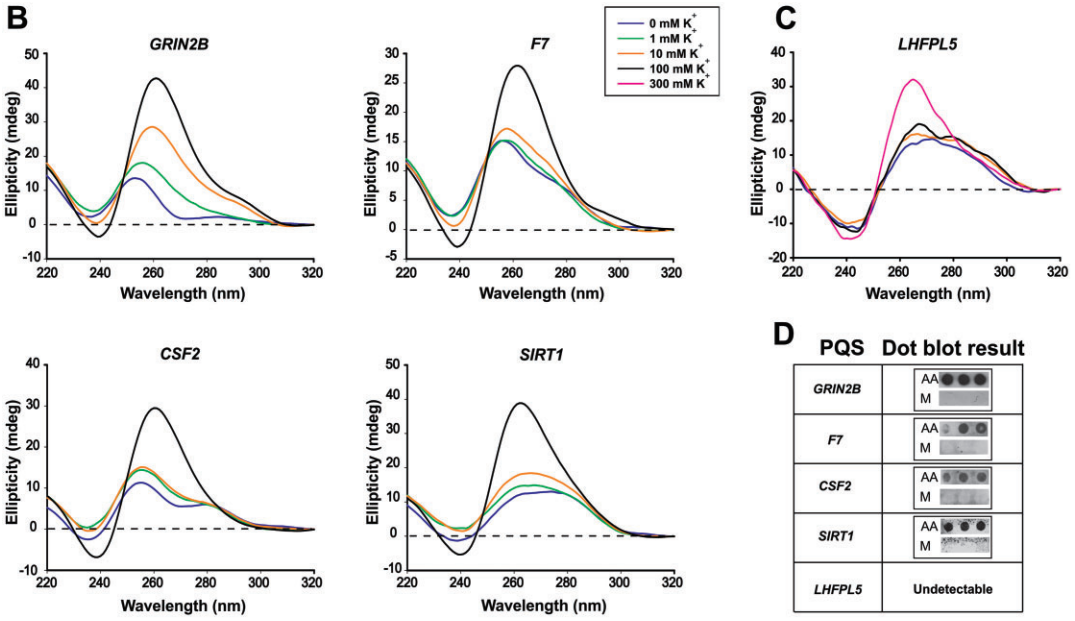


Figure 2. Evidence of *in vitro* G4 folding of PQSs containing the selected G4-Vars. **(A)** Main features of the sequences containing the PQSs including the five selected G4-Vars. Sequences from the sub-group A2 correspond to AA-sequences, while the sequence from the sub-group B2 correspond to VA-sequence. PQSs are underlined, G-tracks ≥3 are signalled in bold and the nucleobases involved in the G4-Vars are signalled in red. Distance to TSS informs both the distance to the most proximal downstream transcript and the distance to the most biologically relevant transcript according to Ensembl. **(B)** CD spectra obtained for the AA-sequences containing the PQSs of the four selected G4-Vars from the sub-group A2. **(C)** CD spectra obtained for the VA-sequence containing the PQS of the selected G4-Var from the sub-group B2. Oligonucleotides containing the PQSs were folded in the presence of increasing K^+ concentrations, as indicated (right-bottom corner box). **(D)** Immunodetection with BG4 antibody of G4s formation by PQSs of the selected G4-Vars. AA- and M-sequences containing the PQSs of each selected G4-Var were folded in 100 mM K^+ and dot-blotted for immunodetection using BG4 antibody. VA- and M-sequences were blotted for the G4-Var of the sub-group B2 (*LHFPL5*).

G4 formation based on score, as G4Hunter (51) (which takes into account G-richness and G-skewness of a sequence and provides a G4 propensity score), as well as based on sequence motifs, as Quadron (already used) and QGRS Mapper (40). Supplementary Table S11 shows that the VA-sequences corresponding to G4-Vars from sub-group A2 (*GRIN2B*, *SIRT1*, *CSF2* and *F7*) show lower scores (or not calculated due to the absence of consensus G4 motif) than the AA-sequences. On the contrary, the AA-sequence corresponding to the G4-Var from sub-group B2 (*LHFPL5*) show lower scores (or not calculated due to the absence of consensus G4 motif) than the VA-sequence (in two of the three predictors). As expected, M-sequences showed null scores for all the predictors. These results support the computational strategy for the identification of G4-Vars and their classification into A2 and B2 subgroups.

G4-Vars are located within the most proximal 200 nucleotides upstream the TSS in the cases of *SIRT1*, *F7* and *LHFPL5*, or within the 500 nucleotides most distal of the PPRs upstream the TSS in the cases of *GRIN2B* and *CSF2*. Moreover, G4-Vars in *GRIN2B*, *SIRT1* and *F7* are located on the template strand while G4-Vars in *CSF2* and *LHFPL5* on the coding strand in (Figure 2A and Supplementary Figure S3).

In vitro analysis of G4-forming capability of PQSs containing G4-Vars

As none of the PQSs containing the selected G4-Vars had been previously characterised as G4-forming sequences, we performed dot-blots and CD spectra to assess whether G4 formation is feasible by synthetic single-stranded oligodeoxyribonucleotide sequences (Supplementary Table S1 and Figure 2A) containing each of the four PQSs from sub-group A2 or the PQS from sub-group B2. In all cases, CD spectra show the typical pattern of peaks associated with parallel G4 structures, showing an increase of a positive peak around 264 nm and a negative peak around 240 nm in response to the presence of increasing K^+ concentrations (Figure 2B, C). CD positive peaks reached the maximal intensities in the presence of 100 mM K^+ (Figure 2B), except for the PQS of *LHFPL5*, which reached an evident CD spectrum of parallel G4 at 300 mM K^+, probably indicating less G4 stability or formation propensity (Figure 2C). Dot-blots using BG4 antibody (41) show signals in all the cases except for *LHFPL5* (Figure 2D). As expected, no signal was detected when assessing the corresponding M-sequences (negative controls, Figure 2D). Noticeable, the five genes selected in this work were identified as containing observed G4 sequences (OQs) by a G4 high-throughput

Table 1. Summary of the data for the selected G4-Vars from sub-groups A2 and B2

Sub-group	Variant id	Ancestral allele nucleotide	Variant allele nucleotide	Frequency of variation allele	G4-Var associated gene Ensembl id	G4-Var location human reference genome GRCh38.p12	Gene name	Gene description	Variant consequence	Associated phenotype	Reference
A2	CR0911240	G	T	0.2	ENSG00000273079	Chromosome 12 : 13982130	*GRIN2B*	glutamate ionotropic receptor NMDA type subunit 2B	upstream_gene_variant	Alzheimer disease increased risk	(36)
	CR982413	C	G	< 0.01	ENSG00000057593	Chromosome 13 : 113105748	*F7*	Coagulation factor VII	upstream_gene_variant	Factor VII deficiency	(37)
	CR035513	C	A	0.74	ENSG00000164400	Chromosome 5 : 132073149	*CSF2*	Colony stimulating factor 2	upstream_gene_variant	Reduced severity in atopic dermatitis	(47)
	CR125660	G	A	<0.01	ENSG00000096717	Chromosome 10 : 67884595	*SIRT1*	sirtuin 1	upstream_gene_variant	Myocardial infarction	(48)
B2	rs886061339	A	G	<0.01	ENSG00000197753	Cromosome 6 : 35805336	*LHFPL5*	LHFPL tetraspan subfamily member 5	5_prime_UTR_variant	Deafness, Hearing impairment	(50)

sequencing method to experimentally detect and map G4 structures in the human genome (G4-seq) (8). Also, the PPRs of these genes were identified as containing OQs (5). Moreover, a PQS was identified into the PPR of *SIRT1* in a genome-wide mapping of endogenous G4s by chromatin immunoprecipitation and high-throughput sequencing (G4 ChIP-seq) in the HaCaT cell line (52), indicating G4 formation within this PPR and probably an accessible open chromatin state linked to a high transcriptional activity of this gene. Taken together, data suggest that the novel PQSs containing the five selected G4-Vars are able to fold as stable G4 structures *in vitro*.

In vitro analysis of the effect of G4-Vars on G4s formation

The effect of the G4-Vars on G4s formation was analysed by four different spectroscopic approaches (CD spectra, CD melting, TDS and 1D ^{1}H NMR) and one biochemical method (qPSA). For the spectroscopic approaches, we used the synthetic single-stranded oligodeoxyribonucleotide representing the VA-sequences (for sub-group A2, *GRIN2B*, *SIRT1*, *CSF2* and *F7*) or the AA-sequence (for sub-group B2, *LHFPL5*). Additionally, sequences containing G to A site-specific mutations disrupting PQSs were used as negative controls (Supplementary Table S1 and Figures 3A, E, I and 4A, E). The VA-sequences from sub-group A2 display CD spectra with reduced intensities in the characteristic peaks of parallel G4s (mainly the positive peaks around 264 nm) when compared with the CD spectra for AA-sequences (Figures 3B, F and 4B, F), suggesting that the nucleotide changes reduce the G4s population likely by reducing their stabilities. As expected, a similar behaviour is observed for the AA-sequence from sub-group B2 (*LHFPL5*) when compared with the CD spectrum for the VA-sequence (Figure 3J). Regardless of the K^{+} concentration, M-sequences show CD spectra with peaks not characteristic of G4s (Figures 3 and 4). G4s thermal stabilities calculated by CD melting show that melting temperatures (Tm) obtained for the VA-sequences from sub-group A2 are lower than those obtained for the AA-sequences, mainly for *CSF2* and *F7* (Supplementary Figure S4). On the contrary, as expected for a G4-Var from sub-group B2, the Tm obtained for the *LHFPL5* AA-sequence is lower than that obtained for the VA-sequence (Supplementary Figure S4). In agreement, TDS spectra for the AA-sequences from sub-group A2 and the VA-sequence from sub-group B2 show the typical G4 signature with two positive peaks around 243 and 273 nm and a negative peak at 295 nm (Supplementary Figure S5). Except for *SIRT1*, the VA-sequences from sub-group A2 and the AA-sequence from sub-group B2 show TDS spectral changes consistent with G4 structural changes. In all cases, the M-sequences show TDS spectra not characteristic of G4. 1D ^{1}H NMR showed a group of defined imino protons signals around 11–12 ppm for the five selected PQSs containing the AA-sequences in the case of PQSs from sub-group A2 or the VA-sequence in the case of the PQS from sub-group B2 (Figures 3C, G, K and 4C, G), confirming the presence of Hoogsteen bonds and G4 structures (53). All the VA-sequences from sub-group A2 and the AA-sequence from sub-group B2 display qualitative (Figures 3C, G, K and 4C, G) and quantitative (Supplementary Figure S6) differences in 1D ^{1}H NMR spectra when compared with their corresponding AA- and VA-sequences. Indeed, although spectra show G4 signatures, the intensity of signals are lower, suggesting that the G4-Vars

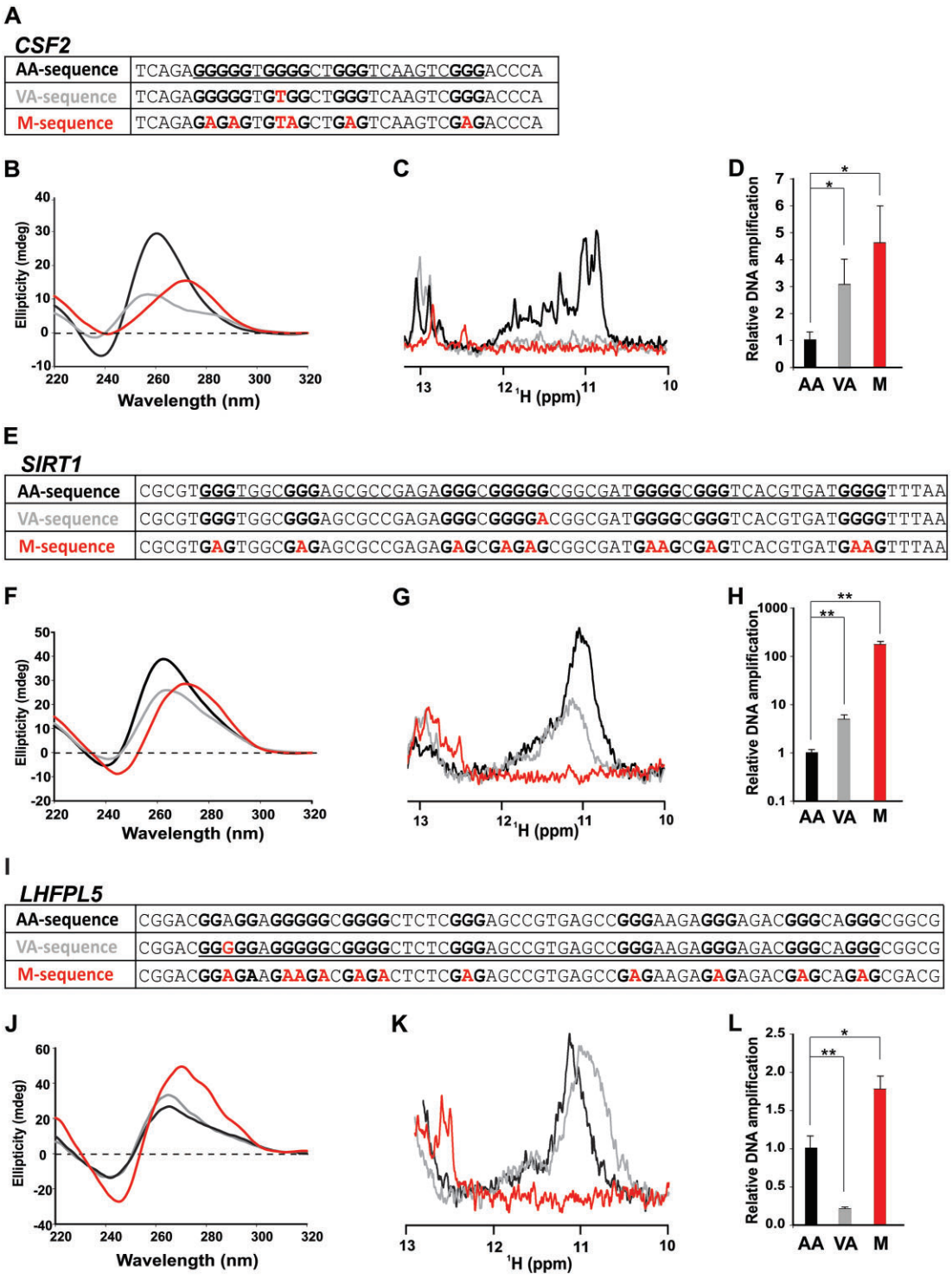


Figure 3. Analysis *in vitro* of G4-Vars consequences in G4 formation by PQSs of *CSF2*, *SIRT1* (sub-group A2) and *LHFPL5* (sub-group B2). **(A)**, **(E)** and **(I)** show tables with the AA-, VA- and M-sequences containing the PQSs of each selected G4-Var. PQSs are underlined, G-tracks ≥ 3 are signalled in bold and the nucleobases involved in the G4-Vars are signalled in red. **(B)**, **(F)** and **(J)** show the CD spectra performed for oligonucleotides folded in the presence of 100 mM K^+ **(B and F)** or 300 mM K^+ **(J)**. **(C)**, **(G)** and **(K)** show the imino region of the 1D ^{1}H NMR spectra obtained for each oligonucleotide sequence folded in the presence of the same K^+ concentration used for CD. **(D)**, **(H)** and **(L)** qPSAs results showing the relative DNA amplification (bars represent the mean of triplicate experiments) relativized to mean results obtained for AA-templates. Error bars correspond to standard deviation (SD). Results for AA-, VA- and M-sequences are represented with black, grey and red colours, respectively. $*P < 0.05$, $**P < 0.01$, Student's *t*-test.

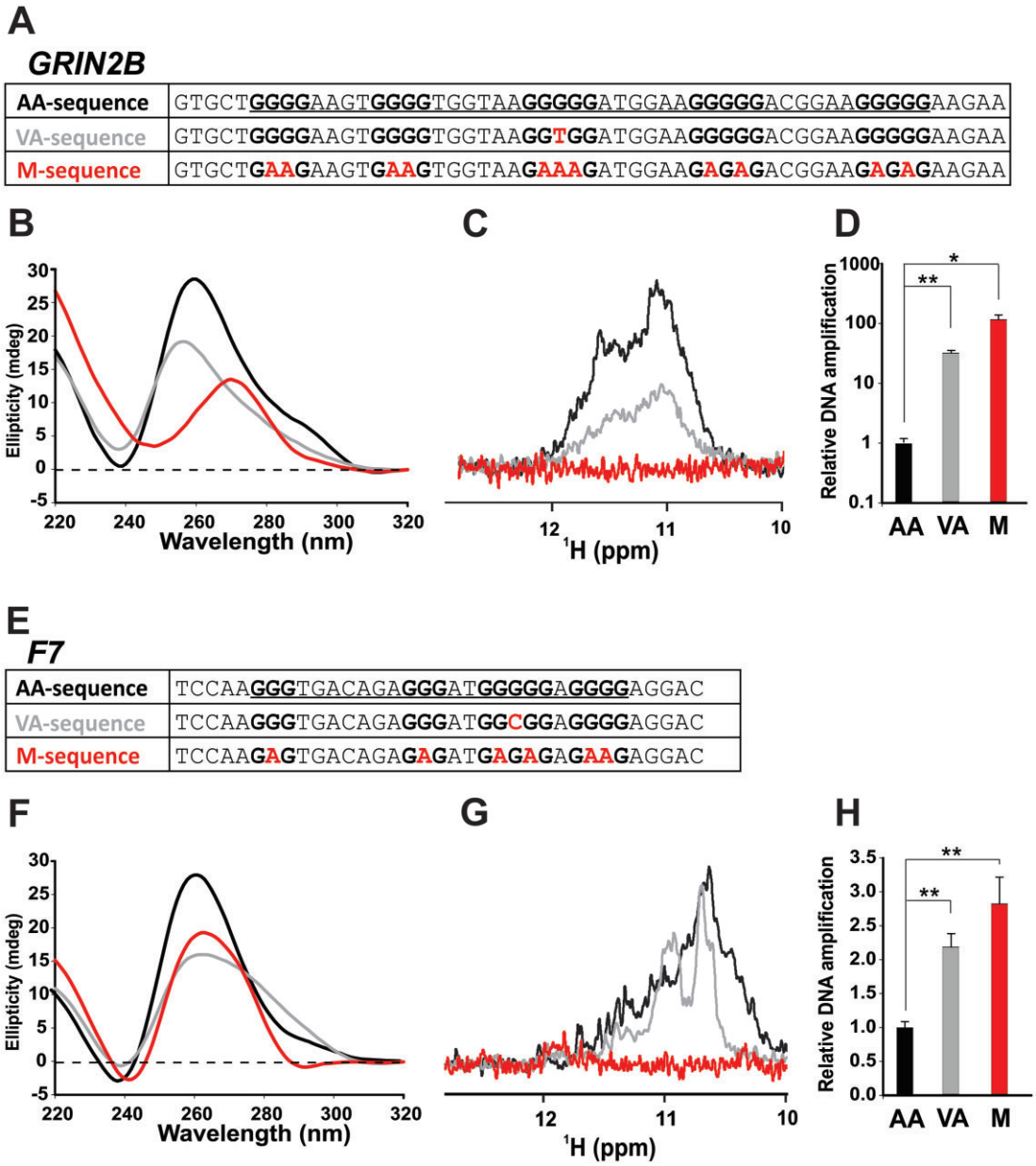


Figure 4. Analysis *in vitro* of G4-Vars consequences in G4 formation by PQSs of *GRIN2B* and *F7* (sub-group A2). **(A)** and **(E)** show tables with the AA-, VA- and M-sequences containing the PQSs of each selected G4-Var. PQSs are underlined, G-tracks ≥ 3 are signalled in bold and the nucleobases involved in the G4-Vars are signalled in red. **(B)** and **(F)** show the CD spectra performed for oligonucleotides folded in presence of 10 mM K^+ **(B)** or 100 mM K^+ **(F)**. **(C)** and **(G)** show the imino region of the 1D ^{1}H NMR spectra obtained for each oligonucleotide sequence folded in the presence of the same K^+ concentration used for CD. **(D)** and **(H)** qPSAs results showing the relative DNA amplification (bars represent the mean of triplicate experiments) relativized to mean results obtained for AA-templates. Error bars correspond to standard deviation (SD). Results for AA-, VA- and M-sequences are represented with black, grey and red colours, respectively. $*P < 0.05$, $**P < 0.01$, Student's *t*-test.

reduce the G4 population, likely by decreasing G4 stability. 1D ^{1}H NMR spectra for the PQSs within the PPRs of *SIRT1*, *CSF2* and *LHFPL5* show signals around 13 ppm, indicating that these sequences may form structures with Watson–Crick base pairs at some extent (53) that may compete with G4s. As expected, the M-sequences show spectra with no peaks indicating Hoogsteen bonds (i.e. no G4 structures).

The qPSA relies on the amplification of ssDNA templates consisting of a central PQS and flanking sequences for the annealing of primers. G4s formed in the ssDNA templates act as roadblocks to *Taq* polymerase, which stalls replication and leads to a decreased amplification efficiency by qPCR. Therefore, formation of G4s with higher stabilities will boost this effect and reduce the amplification efficiencies, while G4 destabilisation will increase the amplification efficiency (43). Figures 3D, H, L and 4D, H show that qPSA amplification efficiencies for the VA-sequences from sub-group A2 and the AA-sequence from sub-group B2 are higher than the qPSA amplification efficiencies for their AA-sequences and VA-sequence, respectively, confirming that G4-Vars disrupting PQSs reduce the G4s population likely by reducing their stabilities. In all cases, the M-sequences show higher qPSA amplification efficiencies indicating no G4 formation and maximal amplification.

Pyridostatin (PDS) is a widely used G4 stabiliser molecule (5,8,54). CD spectra performed on sequences pre-incubated with PDS show that the sequences with disrupted PQSs are more responsive to increasing PDS concentrations than their counterparts containing PQSs (Supplementary Figure S7). qPSAs performed in the presence of PDS (Supplementary Figure S8) show that the VA-sequences for *GRIN2B*, *CSF2* and *SIRT1* are more susceptible to be stabilised by PDS than the AA sequences. These results suggest that G4-stabilising ligands revert the effect of the G4-Vars on G4s stabilities.

Collectively, *in vitro* results show that the selected G4-Vars affect G4s formation and stabilities, and suggest that this can be reverted by the presence of G4-specific ligands. These facts may account for functional effects in transcriptional gene expression control.

Analysis of G4-Vars consequences on transcriptional regulation in a cellular context

The next key question was whether the G4-Vars have consequences on transcriptional activity. To address the effect of G4s and G4-Vars in the transcription process in a cellular context, we constructed luciferase reporter plasmids for the five G4-Vars by cloning each PQS, as well as their variant and mutated versions, upstream of the SV40 basal promoter into the pGL3-promoter vector in the same DNA-strand (template or coding) in which they are found in the human genome (Supplementary Figure S9A). The formation of G4 structures in the plasmids was evaluated using qPSA, employing them as templates (Supplementary Figure S9A). G4 formation was evaluated both in plasmids previous to the transfections (pre-transfection) and in those recovered from transiently transfected HEK-293 cells (post-transfection). Although qPSA is not a G4-specific method, we reasoned that if G4s were indeed formed in the plasmid context they may act as roadblocks to *Taq* polymerase, thus reducing the amplification efficiency (55). In all cases, regardless of whether the plasmids were evaluated pre- or post-transfection, the presence of the corresponding PQSs cause a reduction in the amplification efficiency compared to those obtained for the empty pGL3-promoter vector (Supplementary Figure S9). In addition, mutated PQSs (unable to fold *in vitro* as G4s) show higher amplification efficiencies than those obtained for the plasmids with cloned PQSs and similar to those obtained for the empty pGL3-promoter vector. These results indicate that G4s can form in plasmids, both pre- and post-transfection, and that the mutated versions disrupt G4s formation.

Transiently transfected HEK-293 cells were then used for measuring the firefly luciferase activity as a transcriptional reporter. Luciferase activities measured in extracts from cells transfected with the plasmids containing the *GRIN2B*- and *SIRT1*-AA-sequences are significantly lower than those measured in extracts from cells transfected with the empty pGL3-promoter vector or the respective *GRIN2B*- and *SIRT1*-M-sequences (Supplementary Figures S10A and D). These results suggest that the *GRIN2B*- and *SIRT1*-AA-sequences, capable of folding as G4s, repress transcription of the reporter gene. On the other hand, luciferase activities measured in extracts from cells transfected with the plasmid containing the *F7*- and *CSF2*-AA-sequences, as well as *LHFPL5*-VA-sequence, are significantly higher than those in extracts from cells transfected with either the empty pGL3-promoter vector or the plasmid containing the respective *F7*-, *CSF2*- and *LHFPL5*-M-sequences (Supplementary Figures S10B, C and E). These results suggest that the *F7*- and *CSF2*-AA-sequences, as well as *LHFPL5*-VA-sequence, all capable of folding as G4, enhance transcription of the reporter gene.

Once we confirmed that the cloned PQSs in the plasmids fold as G4s and that these G4s affect the transcription of a reporter gene in transiently transfected HEK-293 cells, we next used qPSAs and LRAs for comparing the G4 formation capability and transcriptional effect, respectively, of the AA- and VA-sequences corresponding to each G4-Var. With the exception of *LHFPL5* analysed post-transfection, which exhibits lower amplification in the AA-sequence than in the VA-sequence (Figure 5I, right panel), the impairment of G4 folding by the G4-Vars, both in pre- or post-transfection assays, leads to higher amplification efficiencies in qPSAs (Figures 5A, C, E, G and I), probably due to destabilisation or loosening of the G4 structures. Results from LRAs using the plasmids containing the *GRIN2B*- and *SIRT1*-VA-sequences showed that the repressive effect on transcription observed for *GRIN2B*- and *SIRT1*-AA-sequences is released by the G4-Vars (Figures 5B and H). On the other hand, when using the plasmids containing the *F7*- and *CSF2*-VA-sequences, as well as *LHFPL5*-AA-sequence, LRAs show that the transcriptional enhancer effect observed for *F7*- and *CSF2*-AA-sequences, as well as *LHFPL5*-VA-sequence, is reduced by the G4-Vars (Figures 5D, F and J). Regardless of the type of effect generated by the G4-Vars in each case, it is worth noticing that in the analysed cases, the gain/loss or structural change of a G4 affects gene expression transcription in a cellular context.

The SNV causing the G4-Var identified within the PPR of *GRIN2B* gene (CR0911240 retrieved from HGMD-PUBLIC) had been previously found in a systematic screening of the PPR of *GRIN2B* in the North Chinese population for detecting possible genetic variants genetically and functionally associated with sporadic Alzheimer's disease (SAD). Genetic analysis further confirmed that homozygosis in AA (defined as wild-type haplotype) increase the risk for SAD, even within subjects without APOE ε4 allele (considered a risk-factor for AD and for early age of disease onset). In coincidence with our results, a fragment of the PPR of *GRIN2B* containing the AA-sequence of the SNV causing the G4-Var identified in our study led to lower transcriptional levels of a reporter gene transiently transfected both in human neuroblastoma (SH-SY5Y) and epithelial carcinoma (HeLa) cell lines compared to the VA-sequence (36). On the other hand, the SNV causing the G4-Var identified within the PPR of *F7* gene (CR982413 retrieved from HGMD-PUBLIC) had been previously identified as a mutation responsible for a severe bleeding familiar disorder (factor VII deficiency, hemarthrosis and chronic arthropathy) in a patient with homozygous VA (defined as mutant-type haplotype). In agreement with our results, the *F7* VA-sequence present in two different fragments of the *F7*- PPR led to a reduction in the expression of transiently transfected reporter genes in the human hepatocellular carcinoma (HEPG2) cell line compared to the AA-sequence (37,56).

Results from LRA of transiently transfected plasmids do not reflect the real chromatin conditions and endogenous transcriptional regulation. However, they represent an experimental approach for evidencing the effect of these novel G4s and G4-Vars in the transcription process in a cellular context. To go deeper in this characterization, we performed RT-qPCR analyses to assess the endogenous expression of the genes under study in several cell lines. We identified HEPG2 as a

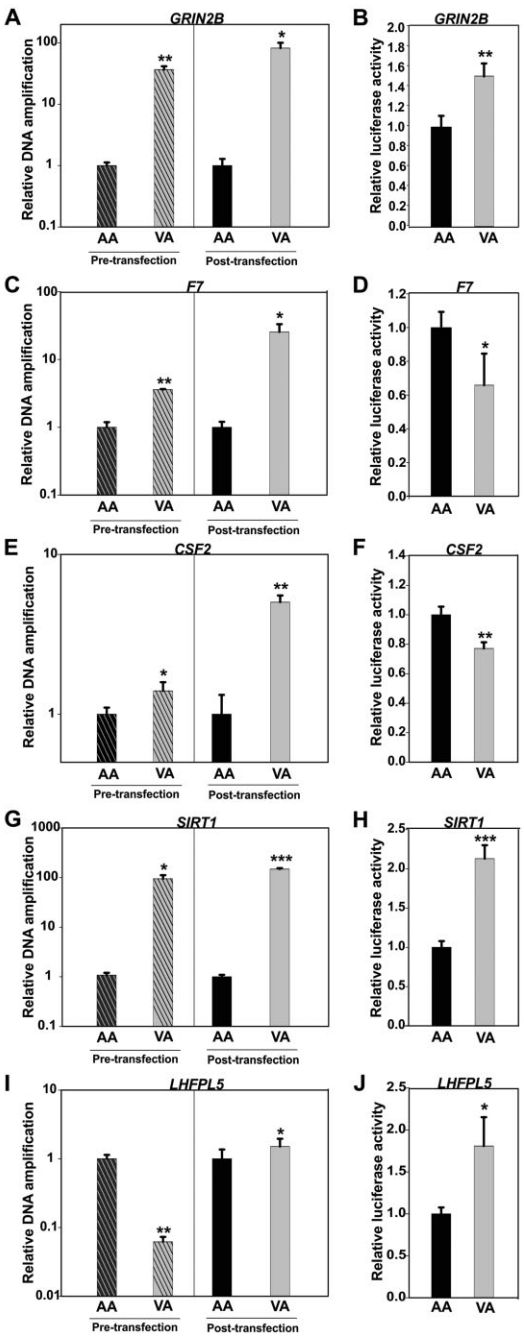


Figure 5. G4-Vars consequences in G4 formation by PQSs cloned in the reporter plasmids and in transcriptional regulation. **(A)**, **(C), (E), (G)** and **(I)** show qPSA results using as templates the plasmids containing AA- and VA-sequences of each selected G4-Var inserted upstream the basal promoter SV40 of pGL3-promoter vector previous to the transfections (pre-transfection, left, striped bars) or recovered from transiently transfected HEK-293 cells (post-transfection, right, empty bars). Results are expressed as the DNA amplification (bars represent the mean of triplicate experiments) relativized to the mean values obtained for the corresponding AA-sequences. **(B)**, **(D), (F), (H)** and **(J)** show LRA results performed in HEK-293 cells transfected with pGL3-promoter vector containing AA- and VA-sequences of each selected G4-Var inserted upstream the basal promoter SV40 of pGL3-promoter vector. Results are expressed as the luciferase activity (mean of three independent experiments) relativized to the mean values obtained for the corresponding AA-sequences. Error bars correspond to standard deviation (SD). Results for AA- and VA-sequences are represented with black and grey colours, respectively. $^{*}P < 0.05$, $^{**}P < 0.01$, $^{***}P < 0.001$, Student's *t*-test.

suitable cell line for modelling *F7* physiological expression, since it expresses detectable levels of *F7* mRNA, it is derived from a human hepatocellular carcinoma, and exerts several features of differentiated hepatocytes, which naturally express *F7*. In addition, the HEPG2 cell line contains the AA-sequence of *F7* PQS, as verified by PCR amplification and sequencing (not shown). Therefore, we evaluated the effect of treating the HEPG2 cell line with specific G4 ligands on the endogenous expression of the *F7* gene. Treatment of HEPG2 cells with TMPyP4, a ligand that according to our *in vitro* results induces the destabilization of the G4 formed by *F7* PQS, leads to a decrease in endogenous *F7* transcriptional expression (Supplementary Figure S11), resembling the effect produced by the G4-Var. Assuming that TMPyP4 maintains its effect on the G4 present in the *F7* PPR in a cellular context, these results are in agreement with results gathered both *in vitro* and in LRAs, suggesting that the endogenous transcription of *F7* is enhanced by the G4 located into its PPR, while repressed by the G4 destabilization.

The combination of bibliographic data with the results in the cellular context presented here suggest that the selected G4-Vars, within the framework of the PPRs, induce alterations in the G4s folding and stability. This leads us to infer that these variations might contribute to changes in the expression of genes relevant for the onset or susceptibility of human genetic diseases.

Discussion

Genetic variations throughout the genome found overlapping with PQSs and affecting G4s formation (G4-Vars) can be envisaged as the 'G4-variome'. The contribution of the G4-variome to the functional diversity of the human genome has been brought to light during the last decade. Indeed, several studies focused on RNA G4s have shown that G4-Vars may perturb the translation, stability and localization of mRNAs (57–60), and affect the microRNA biogenesis (61). The potential impact of the G4-Vars on DNA biology began to emerge by the identification G4-Vars in genome-wide studies (25,27,28,62–64) and by the characterisation of the role of specific G4-Vars in the transcriptional regulation of particular genes (34,35). A pioneer genome-wide analysis of SNVs in human PQSs was performed emphasising on G4-Vars occurrences both in genes and their regulatory sequences. Data suggested that disruptive variations in G-tracks of PQSs are less frequent than neutral variations in loops (62). Then, a merged analysis of genotype information, SNV data and gene expression profiles assessed whether the difference in the expression of particular genes is associated with the G4-Vars. Bioinformatics and experimental results suggested both a relative selection bias against alteration of PQSs and a significant role of G4-Vars in gene expression variations among individuals (27). Based on these findings, G4-Vars were suggested as *cis* quantitative trait loci associated with expression (*cis*-eQTLs, or loci responsible for quantitative alteration in gene expression) capable of causing substantial alteration of promoters activities (28). A more recent work has reported the existence of more than 5 million pG4-Vars causing gains/losses or structural conversions of PQSs within the human genome, most of them enriched near the TSSs and mainly overlapping with transcription factor–binding sites (TFBSs) and enhancers. This finding envisaged important putative implications of the G4-variome on gene activity and positioned the G4-Vars as

a novel category of targets for personalised health risk assessment and drug development (25). In line with these concepts, recent integrative genomic analyses of G4s present in promoters revealed they are enriched with *cis*-eQTL variants, histone modifications and transcription factors (TFs) binding sites (31), reinforcing the biological relevance of the G4s in promoters regulating gene expression and the impact of the G4-variome on this function. Moreover, it has been proposed that G4-Vars are also relevant for gene expression variations in organisms of commercial interest, such as barley (63) and cattle (64), thus potentially linking G4-Vars to traits of agricultural and livestock importance.

G4s, along with other non-B DNA structures, have emerged as major contributors to genome-wide SNVs by disrupting replication and transcription (65–68), thus leading to genetic diversity (29–32,69,70) and probably causing at least part of the G4-Vars. However, recent insights question the role of G4s as major instigators of mutagenesis and instead, these structures appear to be mildly mutagenic but potent causative of recurrent sequencing errors (69,71). This calls for caution in low-read-depth sequencing studies and rare variant analysis, which should be studied by a comprehensive strategy involving a combination of several independent sequencing technologies, increased read depth and strict quality filters (69,71).

Although several works have pointed out a link between the G4-variome and the appearance of diseases, a comprehensive search intended for identifying pathogenic G4-Vars with possible consequences on gene transcription has not yet been carried out. In this work, we developed a novel strategy aimed at identifying and testing disease-associated G4-Vars located into PPRs likely affecting transcriptional gene activity. It should be noted that our original computational pipeline is not restricted solely to the goal of this work, and can also be adapted to identify G4-Vars affecting other steps of gene expression or even other pathways in which the G4s are involved.

We have set high stringent criteria for the selection of the G4-Vars to be assessed, being aware that this stringency may have circumvented some biologically relevant G4-Vars. We restricted the analysis to those G4-Vars with high chances of impairing or favouring G4s formation due to disruption or promotion of a PQS (groups A and B); i.e. G4-Vars that mainly affect G-tracks, thus affecting the nucleotides defining the core structure of G4s. Recent evidence revealing high selective pressure on PQSs forming stable G4s within promoters reinforced our selection. Particularly, G-tracks within these PQSs displayed less tolerance for changes compared to loops and flanking regions (31,32), indicating that G-tracks are pivotal elements for G4s structure and biological function. Moreover, we selected by Quadron algorithm (39) the PQSs with the highest propensity to form G4s *in vitro* (subgroups A2 and B2) with the extended canonical consensus definition. This consensus includes 65% of the observed G4s by G4-Seq (39), but do not consider the non-canonical sequences (that contain bulges in the G-tracks or loops longer than 12 nucleotides), which would provide additional complexity to the G4-variome scenario. In addition, we left behind G4-Vars probably containing SNVs within the loops, the flanking regions, or even the G-tracks as long as they keep the presence of the PQSs. These G4-Vars represent the majority of the identified pG4-Vars (83%, or 15712 ids out of a total of 18924) and, even conserving a canonical PQS, could alter the G4s stability and/or topology, thus probably accounting for relevant changes in transcriptional levels of genes regulated by G4s. Therefore, we consider this work as a start point for discovering the implications of the G4-variome in the onset of human genetic diseases or even in the possible treatment, diagnosis, and/or prevention of these diseases.

The five G4-Vars assessed here comprise a variety of cases including G4 disruption (four G4-Vars from sub-group A2) and promotion (one G4-Var from sub-group B2), with PQSs located on the template (three G4-Vars) and coding (two G4-Vars) strands in respect to transcription, and at different distances from TSSs within the PPRs. The five G4-Vars had been formerly associated with pathological phenotypes of diverse nature, including neurological and cardiac diseases, atopic dermatitis, coagulopathy and deafness, but had not been related with G4 structures. Importantly, by means of a battery of biophysical and biochemical methods, we both confirmed that these G4-Vars are located into PQSs able to fold as G4s not previously reported and verified the expected effect of them on the G4 formation (i.e. G4 disruption or promotion), thus pointing out the success and robustness of the computational strategy. In the cases of G4-Vars found in *GRIN2B* and *F7* promoters, results gathered in cultured cells recapitulated reports by other groups describing the effect of these SNVs on transcription of transiently transfected reporter genes in other cell types relevant to the studied disease (36,37,56). Reinforcing this notion, the action of a G4-specific ligand on the G4 located in the *F7* PPR suggests that the underlying molecular mechanism responsible for the transcriptional effect generated by the SNV could involve the formation of a G4 structure. This finding allows to speculate that G4-Vars might account for changes in the expression of other genes relevant for the onset of and/or predisposition for other pathologies, such as Alzheimer's disease (in the case of *GRIN2B*), atopic dermatitis (*CSF2*), myocardial infarction (*SIRT1*) and deafness (*LHFPL5*). It is important to note that, while different G4 predictors may be suitable tools to foresee changes in the folding and stabilities of the G4s *in vitro*, the consequences of the G4-Vars on the transcriptional control in a cellular context are not linearly predictable from the computational data presented here. In consequence, each G4-Var needs to be evaluated experimentally regarding its effect on G4 formation and its role on transcriptional activity in a relevant biological context, so as to assess their real significance in genes expression regulation and disease.

Although the influence of G4 structures *per se* on transcription activity has been widely described (2,68), the combined effects of DNA-binding proteins acting as TFs together with G4 folding should also be considered, as G4s may act as binding hubs for many different TFs influencing transcription (26,72,73). Indeed, the G4-Var in the *GRIN2B* PPR overlaps with a putative binding site for the zinc finger ras-responsive element binding protein (RREB), reported as a transcriptional repressor. Therefore, it was suggested that the AA-sequence allows the binding of RREB, thus repressing transcription, while the VA-sequence impairs RREB binding, thus promoting transcriptional activity (36). Regarding the G4-Var found in the *F7* PPR, the VA-sequence prevents the binding of SP1 and other nuclear proteins leading to transcriptional repression (37,56). The link between PQSs and the occurrence of the SP1-binding sequences has been extensively characterised (74–76), and SP1 has been described as a G4 binding protein (26,77–79). Alterations in G4 folding may adversely impact on SP1 binding affinity and, consequently, on its functions as a TF (78,80).

In line with this, a predictive analysis of TFBSs performed on AA- and VA-sequences for the five G4-Vars assessed here shows that nucleotide variations modify the predicted TFBSs for several TFs in both the AA- and the VA-sequences (Supplementary Table S12 and Supplementary File S5). Among the TFBS differentially predicted between AA- and VA-sequences, several of them are G4-binding proteins or TFs formerly described as able to bind to sequences overlapping or close to PQSs; e. g., MAZ, E2F1, p53 and WT1. Therefore, it is tempting to speculate that the binding of these TFs is modified by changes in G4s folding or stabilities caused by G4-Vars, thus impacting on transcriptional activity.

This work is a pioneer exploration of the G4-variome carried out focusing on the transcription of genes for which SNVs have been related to human diseases. The identification of G4-Vars related to human pathologies is not only useful for disease diagnosis, but may also indicate the relevance of a particular G4 structure and thus serve as a target for drug development and design of disease-specific treatments. The effect of G4-Vars promoting or disrupting G4 formation could be reverted by treatments with specific ligand molecules destabilising or stabilising an affected G4 structure, thus restoring the gene expression levels. Therefore, scientific efforts should be focused on deepening in the discovery and design of sequence-specific G4 ligands, thus contributing to the design of personalised therapeutic approaches.

Data availability

All data obtained and presented in this work are available in the manuscript and as Supplementary Tables and Files. Quantitative PCR comply with the MIQE Guidelines as detailed in Materials and Methods section. Synthetic oligonucleotides sequences and source are detailed in Materials and Methods section and in Supplementary Table 1. Custom scripts are available as Supplementary Files 1 and 2. Complete data obtained from bioinformatics approaches are available as Supplementary Tables 2 to 12 and Supplementary Files 3 to 5.

Supplementary data

Supplementary Data are available at NAR Online.

Acknowledgements

We are thankful to Dolores Campos for excellent cell culture assistance and María Bucci Muñoz, from Instituto de Fisiología Experimental (IFISE)—CONICET-UNR, for providing the HEPG2 cell line.

Author contributions: P.A. and N.B.C. performed the conceptualization and design of the work. M.G. and E.M. developed the bioinformatic pipeline to identify disease-related G4-Vars, and computationally characterised their related genes. E.J.P. obtained the variation databases and sequences data and A.L. performed the analysis of transcription factors binding sites. A.L. performed most experiments with the assistance of E.J.P. A.B. performed and analysed NMR spectroscopies. P.A. and N.B.C. were responsible for supervision, funding acquisition, project administration and obtaining of resources. P.A. conducted the original draft writing and visualization, while P.A., A.L., E.J.P. and N.B.C. performed the major writing review and editing, assisted by M.G., E.M. and A.B. All authors have read and agreed to the published version of the manuscript.

Funding

Agencia Nacional de Promoción Científica y Tecnológica [PICT 2016-0671 to N.B.C., PICT 2017-0976 and PICT 2019-1662 to P.A.]; Consejo Nacional de Investigaciones Científicas y Técnicas [2015-0170 to N.B.C.]; Universidad Nacional de Rosario [BIO573 to P.A.].

Conflict of interest statement

None declared.

References

1. Wells,R.D. (2007) Non-B DNA conformations, mutagenesis and disease. *Trends Biochem. Sci*, **32**, 271–278.
2. Armas,P., David,A. and Calcaterra,N.B. (2017) Transcriptional control by G-quadruplexes: in vivo roles and perspectives for specific intervention. *Transcription*, **8**, 21–25.
3. Bartas,M., Čutová,M., Brázda,V., Kaura,P., Šťastný,J., Kolomazník,J., Coufal,J., Goswami,P., Červeň,J. and Pečinka,P. (2019) The presence and localization of G-quadruplex forming sequences in the domain of bacteria. *Molecules*, **24**, 1711.
4. Brázda,V., Luo,Y., Bartas,M., Kaura,P., Porubiaková,O., Šťastný,J., Pečinka,P., Verga,D., Da Cunha,V., Takahashi,T.S., *et al.* (2020) G-quadruplexes in the archaea domain. *Biomolecules*, **10**, 1349.
5. Marsico,G., Chambers,V.S., Sahakyan,A.B., McCauley,P., Boutell,J.M., Antonio,M.D. and Balasubramanian,S. (2019) Whole genome experimental maps of DNA G-quadruplexes in multiple species. *Nucleic Acids Res.*, **47**, 3862–3874.
6. Ruggiero,E., Zanin,I., Terreri,M. and Richter,S.N. (2021) G-quadruplex targeting in the fight against viruses: an update. *Int. J. Mol. Sci.*, **22**, 10984.
7. Huppert,J.L. (2005) Prevalence of quadruplexes in the human genome. *Nucleic Acids Res.*, **33**, 2908–2916.
8. Chambers,V.S., Marsico,G., Boutell,J.M., Di Antonio,M., Smith,G.P. and Balasubramanian,S. (2015) High-throughput sequencing of DNA G-quadruplex structures in the human genome. *Nat. Biotechnol.*, **33**, 877–881.
9. Lane,A.N., Chaires,J.B., Gray,R.D. and Trent,J.O. (2008) Stability and kinetics of G-quadruplex structures. *Nucleic Acids Res.*, **36**, 5482–5515.
10. Neidle,S. and Parkinson,G.N. (2003) The structure of telomeric DNA. *Curr. Opin. Struct. Biol.*, **13**, 275–283.
11. Armas,P. and Calcaterra,N.B. (2018) G-quadruplex in animal development: contribution to gene expression and genomic heterogeneity. *Mech. Dev.*, **154**, 64–72.
12. David,A.P., Margarit,E., Domizi,P., Banchio,C., Armas,P. and Calcaterra,N.B. (2016) G-quadruplexes as novel cis-elements controlling transcription during embryonic development. *Nucleic Acids Res.*, **44**, 4163–4173.
13. Nakanishi,C. and Seimiya,H. (2020) G-quadruplex in cancer biology and drug discovery. *Biochem. Biophys. Res. Commun.*, **531**, 45–50.
14. 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.
15. Zahler,A.M., Williamson,J.R., Cech,T.R. and Prescott,D.M. (1991) Inhibition of telomerase by G-quartet DNA structures. *Nature*, **350**, 718–720.
16. Paudel,B.P., Moye,A.L., Abou Assi,H., El-Khoury,R., Cohen,S.B., Holien,J.K., Birrento,M.L., Samosorn,S., Intharapichai,K.,

Tomlinson,C.G., *et al.* (2020) A mechanism for the extension and unfolding of parallel telomeric G-quadruplexes by human telomerase at single-molecule resolution. *eLife*, **9**, e56428.
17. Hirashima,K. and Seimiya,H. (2015) Telomeric repeat-containing RNA/G-quadruplex-forming sequences cause genome-wide alteration of gene expression in human cancer cells in vivo. *Nucleic Acids Res.*, **43**, 2022–2032.
18. Paeschke,K., Bochman,M.L., Daniela Garcia,P., Cejka,P., Friedman,K.L., Kowalczykowski,S.C. and Zakian,V.A. (2013) Pif1 family helicases suppress genome instability at G-quadruplex motifs. *Nature*, **497**, 458–462.
19. De Magis,A., Manzo,S.G., Russo,M., Marinello,J., Morigi,R., Sordet,O. and Capranico,G. (2019) DNA damage and genome instability by G-quadruplex ligands are mediated by R loops in human cancer cells. *Proc. Natl. Acad. Sci. U.S.A.*, **116**, 816–825.
20. Maizels,N. (2015) G4-associated human diseases. *EMBO Rep.*, **16**, 910–922.
21. Teng,F.Y., Jiang,Z.Z., Guo,M., Tan,X.Z., Chen,F., Xi,X.G. and Xu,Y. (2021) G-quadruplex DNA: a novel target for drug design. *Cell. Mol. Life Sci.*, **78**, 6557–6583.
22. Robinson,J., Raguseo,F., Nuccio,S.P., Liano,D. and Di Antonio,M. (2021) DNA G-quadruplex structures: more than simple roadblocks to transcription?*Nucleic Acids Res.*, **49**, 8419–8431.
23. Li,B. and Ritchie,M.D. (2021) From GWAS to gene: transcriptome-wide association studies and other methods to functionally understand GWAS discoveries. *Front. Genet.*, **12**, 713230.
24. Lappalainen,T. and MacArthur,D.G. (2021) From variant to function in human disease genetics. *Science*, **373**, 1464–1468.
25. Gong,J.Y., Wen,C.J., Tang,M.L., Duan,R.F., Chen,J.N., Zhang,J.Y., Zheng,K.W., He,Y.D., Hao,Y.H., Yu,Q., *et al.* (2021) G-quadruplex structural variations in human genome associated with single-nucleotide variations and their impact on gene activity. *Proc. Natl. Acad. Sci. U.S.A.*, **118**, e2013230118.
26. Lago,S., Nadai,M., Cernilogar,F.M., Kazerani,M., Domíniguez Moreno,H., Schotta,G. and Richter,S.N. (2021) Promoter G-quadruplexes and transcription factors cooperate to shape the cell type-specific transcriptome. *Nat. Commun.*, **12**, 3885.
27. Baral,A., Kumar,P., Halder,R., Mani,P., Yadav,V.K., Singh,A., Das,S.K. and Chowdhury,S. (2012) Quadruplex-single nucleotide polymorphisms (Quad-SNP) influence gene expression difference among individuals. *Nucleic Acids Res.*, **40**, 3800–3811.
28. Baral,A., Kumar,P., Pathak,R. and Chowdhury,S. (2013) Emerging trends in G-quadruplex biology-role in epigenetic and evolutionary events. *Mol. Biosyst.*, **9**, 1568–1575.
29. Guiblet,W.M., Cremona,M.A., Harris,R.S., Chen,D., Eckert,K.A., Chiaromonte,F., Huang,Y.F. and Makova,K.D. (2021) Non-B DNA: a major contributor to small- and large-scale variation in nucleotide substitution frequencies across the genome. *Nucleic Acids Res.*, **49**, 1497–1516.
30. Georgakopoulos-Soares,I., Morganella,S., Jain,N., Hemberg,M. and Nik-Zainal,S. (2018) Noncanonical secondary structures arising from non-B DNA motifs are determinants of mutagenesis. *Genome Res.*, **28**, 1264–1271.
31. Li,G., Su,G., Wang,Y., Wang,W., Shi,J., Li,D. and Sui,G. (2023) Integrative genomic analyses of promoter G-quadruplexes reveal their selective constraint and association with gene activation. *Commun. Biol.*, **6**, 625.
32. Makova,K.D. and Weissensteiner,M.H. (2023) Noncanonical DNA structures are drivers of genome evolution. *Trends Genet.*, **39**, 109–124.
33. Hänsel-Hertsch,R., Simeone,A., Shea,A., Hui,W.W.I., Zyner,K.G., Marsico,G., Rueda,O.M., Bruna,A., Martin,A., Zhang,X., *et al.* (2020) Landscape of G-quadruplex DNA structural regions in breast cancer. *Nat. Genet.*, **52**, 878–883.
34. Peng,S.-X., Wang,Y.-Y., Zhang,M., Zang,Y.-Y., Wu,D., Pei,J., Li,Y., Dai,J., Guo,X., Luo,X., *et al.* (2021) SNP rs10420324 in the AMPA receptor auxiliary subunit TARP γ-8 regulates the susceptibility to antisocial personality disorder. *Sci. Rep.*, **11**, 11997.
35. Inagaki,H., Ota,S., Nishizawa,H., Miyamura,H., Nakahira,K., Suzuki,M., Nishiyama,S., Kato,T., Yanagihara,I. and Kurahashi,H. (2019) Obstetric complication-associated ANXA5 promoter polymorphisms may affect gene expression via DNA secondary structures. *J. Hum. Genet.*, **64**, 459–466.
36. Jiang,H. and Jia,J. (2009) Association between NR2B subunit gene (GRIN2B) promoter polymorphisms and sporadic Alzheimer's disease in the North Chinese population. *Neurosci. Lett.*, **450**, 356–360.
37. Carew,J.A., Pollak,E.S., High,K.A. and Bauer,K.A. (1998) Severe factor VII deficiency due to a mutation disrupting an Sp1 binding site in the factor VII promoter. *Blood*, **92**, 1639–1645.
38. Hunt,S.E., McLaren,W., Gil,L., Thormann,A., Schuilenburg,H., Sheppard,D., Parton,A., Armean,I.M., Trevanion,S.J., Flicek,P., *et al.* (2018) Ensembl variation resources. *Database (Oxford).*, **2018**, bay119.
39. Sahakyan,A.B., Chambers,V.S., Marsico,G., Santner,T., Di Antonio,M. and Balasubramanian,S. (2017) Machine learning model for sequence-driven DNA G-quadruplex formation. *Sci. Rep.*, **7**, 14535.
40. 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.
41. Biffi,G., Tannahill,D., McCafferty,J. and Balasubramanian,S. (2013) Quantitative visualization of DNA G-quadruplex structures in human cells. *Nat. Chem.*, **5**, 182–186.
42. Hwang,T.L. and Shaka,A.J. (1995) Water suppression that works. Excitation sculpting using arbitrary wave-forms and pulsed-field gradients. *J. Magn. Reson., Ser. A*, **112**, 275–279.
43. Mitteaux,J., Lejault,P., Wojciechowski,F., Joubert,A., Boudon,J., Desbois,N., Gros,C.P., Hudson,R.H.E., Boulé,J.B., Granzhan,A., *et al.* (2021) Identifying G-Quadruplex-DNA-disrupting small molecules. *J. Am. Chem. Soc.*, **143**, 12567–12577.
44. Whale,A.S., De Spiegelaere,W., Trypsteen,W., Nour,A.A., Bae,Y.-K., Benes,V., Burke,D., Cleveland,M., Corbisier,P., Devonshire,A.S., *et al.* (2020) The Digital MIQE Guidelines update: minimum information for publication of quantitative digital PCR experiments for 2020. *Clin. Chem.*, **66**, 1012–1029.
45. Jordan,M., Köhne,C. and Wurm,F.M. (1998) Calcium-phosphate mediated DNA transfer into HEK-293 cells in suspension: control of physicochemical parameters allows transfection in stirred media. Transfection and protein expression in mammalian cells. *Cytotechnology*, **26**, 39–47.
46. Myers,S.J., Yuan,H., Kang,J.-Q., Tan,F.C.K., Traynelis,S.F. and Low,C.-M. (2019) Distinct roles of GRIN2A and GRIN2B variants in neurological conditions. *F1000Res*, **8**, 1940.
47. Rafatpanah,H., Bennett,E., Pravica,V., McCoy,M.J., David,T.J., Hutchinson,I.V. and Arkwright,P.D. (2003) Association between novel GM-CSF gene polymorphisms and the frequency and severity of atopic dermatitis. *J. Allergy Clin. Immunol.*, **112**, 593–598.
48. Cui,Y., Wang,H., Chen,H., Pang,S., Wang,L., Liu,D. and Yan,B. (2012) Genetic analysis of the SIRT1 gene promoter in myocardial infarction. *Biochem. Biophys. Res. Commun.*, **426**, 232–236.
49. Al-Amri,A.H., Al Saegh,A., Al-Mamari,W., El-Asrag,M.E., Al-Kindi,M.N., Al Khabouri,M., Al Wardy,N., Al Lamki,K., Gabr,A., Idris,A., *et al.* (2019) LHFPL5 mutation: a rare cause of non-syndromic autosomal recessive hearing loss. *Eur. J. Med. Genet.*, **62**, 103592.
50. NM_182548.3(LHFPL5):c.-335A>G AND Autosomal recessive nonsyndromic hearing loss 67 - ClinVar - NCBI.
51. Brázda,V., Kolomazník,J., Lýsek,J., Bartas,M., Fojta,M., Šťastný,J. and Mergny,J.-L. (2019) G4Hunter web application: a web server for G-quadruplex prediction. *Bioinformatics*, **35**, 3493–3495.
52. Hänsel-Hertsch,R., Beraldi,D., Lensing,S.V., Marsico,G., Zyner,K., Parry,A., Di Antonio,M., Pike,J., Kimura,H., Narita,M., *et al.* (2016) G-quadruplex structures mark human regulatory chromatin. *Nat. Genet.*, **48**, 1267–1272.

53. Phan,A.T. (2002) Human telomeric DNA: g-quadruplex, i-motif and Watson-Crick double helix. *Nucleic Acids Res.*, **30**, 4618–4625.
54. Rodriguez,R., Müller,S., Yeoman,J.A., Trentesaux,C., Riou,J.F. and Balasubramanian,S. (2008) A novel small molecule that alters shelterin integrity and triggers a DNA-damage response at telomeres. *J. Am. Chem. Soc.*, **130**, 15758–15759.
55. Jamroskovic,J., Obi,I., Movahedi,A., Chand,K., Chorell,E. and Sabouri,N. (2019) Identification of putative G-quadruplex DNA structures in S. pombe genome by quantitative PCR stop assay. *DNA Repair (Amst.)*, **82**, 102678.
56. Barbon,E., Pignani,S., Branchini,A., Bernardi,F., Pinotti,M. and Bovolenta,M. (2016) An engineered tale-transcription factor rescues transcription of factor VII impaired by promoter mutations and enhances its endogenous expression in hepatocytes. *Sci. Rep.*, **6**, 28304.
57. Beaudoin,J.D. and Perreault,J.P. (2010) 5′-UTR G-quadruplex structures acting as translational repressors. *Nucleic Acids Res.*, **38**, 7022–7036.
58. Huijbregts,L., Roze,C., Bonafe,G., Houang,M., Le Bouc,Y., Carel,J.C., Leger,J., Alberti,P. and De Roux,N. (2012) DNA polymorphisms of the KiSS1 3' Untranslated region interfere with the folding of a G-rich sequence into G-quadruplex. *Mol. Cell. Endocrinol.*, **351**, 239–248.
59. Zeraati,M., Moye,A.L., Wong,J.W.H., Perera,D., Cowley,M.J., Christ,D.U., Bryan,T.M. and Dinger,M.E. (2017) Cancer-associated noncoding mutations affect RNA G-quadruplex-mediated regulation of gene expression. *Sci. Rep.*, **7**, 708.
60. Lee,D.S.M., Ghanem,L.R. and Barash,Y. (2020) Integrative analysis reveals RNA G-quadruplexes in UTRs are selectively constrained and enriched for functional associations. *Nat. Commun.*, **11**, 527.
61. Imperatore,J.A., Then,M.L., McDougal,K.B. and Mihailescu,M.R. (2020) Characterization of a G-quadruplex structure in pre-miRNA-1229 and in its Alzheimer's Disease-associated variant rs2291418: implications for miRNA-1229 maturation. *Int. J. Mol. Sci.*, **21**, 767.
62. Nakken,S., Rognes,T. and Hovig,E. (2009) The disruptive positions in human G-quadruplex motifs are less polymorphic and more conserved than their neutral counterparts. *Nucleic Acids Res.*, **37**, 5749–5756.
63. Cagirici,H.B., Budak,H. and Sen,T.Z. (2021) Genome-wide discovery of G-quadruplexes in barley. *Sci. Rep.*, **11**, 7876.
64. Stefos,G.C., Theodorou,G. and Politis,I. (2022) Genomic landscape, polymorphism and possible LINE-associated delivery of G-quadruplex motifs in the bovine genes. *Genomics*, **114**, 110272.
65. Lemmens,B., van Schendel,R. and Tijsterman,M. (2015) Mutagenic consequences of a single G-quadruplex demonstrate mitotic inheritance of DNA replication fork barriers. *Nat. Commun.*, **6**, 8909.
66. Lerner,L.K. and Sale,J.E. (2019) Replication of G quadruplex DNA. *Genes (Basel)*, **10**, 95.
67. Van Wietmarschen,N., Merzouk,S., Halsema,N., Spierings,D.C.J., Guryev,V. and Lansdorp,P.M. (2018) BLM helicase suppresses recombination at G-quadruplex motifs in transcribed genes. *Nat. Commun.*, **9**, 271.
68. Kim,N. (2019) The interplay between G-quadruplex and transcription. *Curr. Med. Chem.*, **26**, 2898–2917.
69. Weissensteiner,M.H., Cremona,M.A., Guiblet,W.M., Stoler,N., Harris,R.S., Cechova,M., Eckert,K.A., Chiaromonte,F., Huang,Y.F. and Makova,K.D. (2023) Accurate sequencing of DNA motifs able to form alternative (non-B) structures. *Genome Res.*, **33**, 907–923.
70. Guiblet,W.M., DeGiorgio,M., Cheng,X., Chiaromonte,F., Eckert,K.A., Huang,Y.F. and Makova,K.D. (2021) Selection and thermostability suggest G-quadruplexes are novel functional elements of the human genome. *Genome Res.*, **31**, 1136–1149.
71. McGinty,R.J. and Sunyaev,S.R. (2023) Revisiting mutagenesis at non-B DNA motifs in the human genome. *Nat. Struct. Mol. Biol.*, **30**, 417–424.
72. Spiegel,J., Cuesta,S.M., Adhikari,S., Hänsel-Hertsch,R., Tannahill,D. and Balasubramanian,S. (2021) G-quadruplexes are transcription factor binding hubs in human chromatin. *Genome Biol.*, **22**, 117.
73. David,A.P., Pipier,A., Pascutti,F., Binolfi,A., Weiner,A.M.J., Challier,E., Heckel,S., Calsou,P., Gomez,D., Calcaterra,N.B., *et al.* (2019) CNBP controls transcription by unfolding DNA G-quadruplex structures. *Nucleic Acids Res.*, **47**, 7901–7913.
74. 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.
75. 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.
76. Kumar,P., Yadav,V.K., Baral,A., Kumar,P., Saha,D. and Chowdhury,S. (2011) Zinc-finger transcription factors are associated with guanine quadruplex motifs in human, chimpanzee, mouse and rat promoters genome-wide. *Nucleic Acids Res.*, **39**, 8005–8016.
77. Raiber,E.A., Kranaster,R., Lam,E., Nikan,M. and Balasubramanian,S. (2012) A non-canonical DNA structure is a binding motif for the transcription factor SP1 in vitro. *Nucleic Acids Res.*, **40**, 1499–1508.
78. Da Ros,S., Nicoletto,G., Rigo,R., Ceschi,S., Zorzan,E., Dacasto,M., Giantin,M. and Sissi,C. (2020) G-quadruplex modulation of SP1 functional binding sites at the KIT proximal promoter. *Int. J. Mol. Sci.*, **22**, 329.
79. Kong,J.-N., Zhang,C., Zhu,Y.-C., Zhong,K., Wang,J., Chu,B.-B. and Yang,G.-Y. (2018) Identification and characterization of G-quadruplex formation within the EP0 promoter of pseudorabies virus. *Sci. Rep.*, **8**, 14029.
80. Tsukakoshi,K., Saito,S., Yoshida,W., Goto,S. and Ikebukuro,K. (2018) CpG methylation changes G-quadruplex structures derived from gene promoters and interaction with VEGF and SP1. *Molecules*, **23**, 944.

Received: August 12, 2022. **Revised:** September 22, 2023. **Editorial Decision:** September 29, 2023. **Accepted:** October 12, 2023