Preferential formation of Z-RNA over intercalated motifs in long noncoding RNA

Uditi Bhatt, Anne Cucchiarini, Yu Luo, et al.

Genome Res. 2024 34: 217-230 originally published online February 14, 2024
Access the most recent version at doi:10.1101/gr.278236.123

Preferential formation of Z-RNA over intercalated motifs in long noncoding RNA

Uditi Bhatt,[1] Anne Cucchiarini,[2] Yu Luo,[2] Cameron W. Evans,[1] Jean-Louis Mergny,[2] K. Swaminathan Iyer,[1] and Nicole M. Smith[1]

[1] *School of Molecular Sciences, The University of Western Australia, Crawley, Western Australia 6009, Australia;*
[2] *Laboratoire d'Optique et Biosciences, École Polytechnique, CNRS, INSERM, Institut Polytechnique de Paris, 91120 Palaiseau, France*

Secondary structure is a principal determinant of lncRNA function, predominantly regarding scaffold formation and interfaces with target molecules. Noncanonical secondary structures that form in nucleic acids have known roles in regulating gene expression and include G-quadruplexes (G4s), intercalated motifs (iMs), and R-loops (RLs). In this paper, we used the computational tools G4-iM Grinder and QmRLFS-finder to predict the formation of each of these structures throughout the lncRNA transcriptome in comparison to protein-coding transcripts. The importance of the predicted structures in lncRNAs in biological contexts was assessed by combining our results with publicly available lncRNA tissue expression data followed by pathway analysis. The formation of predicted G4 (pG4) and iM (piM) structures in select lncRNA sequences was confirmed in vitro using biophysical experiments under near-physiological conditions. We find that the majority of the tested pG4s form highly stable G4 structures, and identify many previously unreported G4s in biologically important lncRNAs. In contrast, none of the piM sequences are able to form iM structures, consistent with the idea that RNA is unable to form stable iMs. Unexpectedly, these C-rich sequences instead form Z-RNA structures, which have not been previously observed in regions containing cytosine repeats and represent an interesting and underexplored target for protein–RNA interactions. Our results highlight the prevalence and potential structure-associated functions of noncanonical secondary structures in lncRNAs, and show G4 and Z-RNA structure formation in many lncRNA sequences for the first time, furthering the understanding of the structure–function relationship in lncRNAs.

[Supplemental material is available for this article.]

The discovery of long noncoding RNAs (lncRNAs) has revolutionized scientific understanding of the genome and has led to an explosion of research into the complex mechanisms that regulate gene expression. Currently, lncRNAs are broadly defined as any RNA >200 nt in length that does not encode for a protein, and it is well understood that lncRNAs display substantial diversity in all aspects, ranging from genomic location to post-transcriptional processing, tissue expression, and function (Chillón and Marcia 2020). Despite the rapid rate of identification of novel lncRNAs, their cellular roles remain poorly characterized. Although lncRNAs generally have poor evolutionary conservation at the nucleotide level, short sequences of lncRNAs have been retained, which was attributed to their potential roles in folding and secondary structure formation (Pegueroles and Gabaldón 2016). In some cases, structural motifs in lncRNAs are conserved, providing further incentive for researchers to explore this area (Johnsson et al. 2014; Tavares et al. 2019). In the past decade, key roles of secondary structure in many lncRNAs have been identified, particularly regarding the formation of scaffolds and functional interfaces with target molecules (Uroda et al. 2019; Simko et al. 2020). The single-stranded nature of RNA allows the formation of various secondary structures such as hairpin loops, triplexes, and double-stranded helices, including A-RNA and Z-RNA, that contribute to the stability and function of the molecule. Despite this, many secondary structures in RNA, particularly noncanonical structures, are significantly less well characterized than their DNA counterparts. In this study, we focus on the formation of G-quadruplex (G4), intercalated-motif (iM), R-loop (RL), and Z-RNA noncanonical secondary structures in lncRNA.

G4s form in certain guanine-rich regions through the self-assembly of guanines into planar quartets, which are stacked into a three-dimensional structure. G4 formation can alter gene expression by inhibiting initiation complex binding at mRNA transcripts, recruiting splicing-associated RNA-binding proteins, or influencing the subcellular localization and stability of RNA molecules. Prior studies have experimentally validated the formation and functional importance of G4s in a few individual lncRNAs; however, little is known in comparison to their counterparts in mRNA and DNA (Biffi et al. 2012; Matsumura et al. 2017; Simko et al. 2020; Bhatt et al. 2021). The roles of G4s in lncRNAs have recently been reviewed (Tassinari et al. 2021). Another tetrameric structure, iMs are found in cytosine-rich regions and form when hemi-protonated cytosine base pairs (C^+···C) intercalate and stack. The existence of RNA iMs in vivo has been contested owing to preferential formation at low pH and notably poorer stability in comparison to DNA iMs, attributed to the presence of the 2′-hydroxyl group in the ribose sugar (Lacroix et al. 1996). These are crucial factors to consider when investigating the relevance of iMs in biological systems. However, immunocytochemistry with an iM structure-specific antibody (iMab) revealed a decrease in iMab foci after RNase treatment, possibly corresponding to the

Corresponding author: nicole.smith@uwa.edu.au
Article published online before print. Article, supplemental material, and publication date are at https://www.genome.org/cgi/doi/10.1101/gr.278236.123.

degradation of RNA iMs, and thus, their formation warrants further investigation (Zeraati et al. 2018).

RLs are triplex structures consisting of an RNA–DNA hybrid and the resulting displaced ssDNA strand. RLs are implicated in DNA replication and repair pathways, and aberrant RL formation can result in DNA damage and replication stress (Crossley et al. 2019; Petermann et al. 2022). Most RLs form during transcription as the nascent RNA anneals with the template DNA strand behind RNA polymerase II (RNA pol II). Many lncRNAs are transcribed by RNA pol II and may be more likely to form RLs owing to the lack of cotranscriptional processing, which actively resolves these structures; however, RLs in lncRNAs have not been thoroughly investigated (Schlackow et al. 2017; Crossley et al. 2019; Yang et al. 2020).

Z-RNA is a left-handed double helix form of DNA and RNA commonly studied in poly-GC repeat sequences in which the alternating *syn–anti* conformation of bases shapes the sugar-phosphate backbone into a zig-zag pattern. The biological roles of Z-RNA have been increasingly studied in the past few years, with particular focus on their interactions with the RNA-editing enzyme adenosine deaminase RNA specific (ADAR; also known as ADAR1) protein (Herbert 2019; Lee et al. 2019; Nakahama and Kawahara 2021). Notably, Z-RNA structures influence the adenine-to-inosine (A-to-I) editing ability of ADAR with downstream effects on autoimmune diseases and encephalopathy (Koeris et al. 2005; de Reuver et al. 2021; Nakahama et al. 2021; Tang et al. 2021; Zillinger and Bartok 2021; Jiao et al. 2022). Despite these advances, little is known regarding Z-RNA formation in different sequences, and the understanding of Z-RNA is lacking in comparison to that of Z-DNA.

Several different transcriptome-wide experimental techniques developed in the past five years have supported the formation of endogenous RNA G4s (rG4s) in human cells. The work of Guo and Bartel (2016) suggested very low rG4 structure formation in cells using dimethylsulfate (DMS) profiling of RNA; however, they did not account for structure folding or unfolding during the experimental procedure. In contrast, Yang et al. (2018) and others showed the prevalent but transient formation of rG4s in the transcriptome by first cross-linking the structures in cells before extracting the RNA (Kwok et al. 2016). Regardless, information on rG4s in lncRNAs is rarely included, instead focusing on poly(A)-enriched or cytoplasmic RNAs (Guo and Bartel 2016; Kwok et al. 2016; Yang et al. 2018; Varshney et al. 2021).

In contrast to rG4s, minimal experimental data on RNA iM formation in human cells currently exist and their formation in vivo is debated, largely owing to their poor stability (Lacroix et al. 1996; Collin and Gehring 1998; Snoussi et al. 2001; Zeraati et al. 2018). The genomic distribution of RLs has been previously determined by multiple groups; however, varying results have been reported depending on the different methods used such as DRIP-seq, RNase H1 ChIP, S9.6 ChIP, and CUT&TAG (Wahba et al. 2016; Chen et al. 2017, 2019; Sanz and Chédin 2019; Wang et al. 2021; Lin et al. 2022). One proposed explanation for these discrepancies is that two distinct classes of RLs exist: relatively short ones (median length, ~200–300 bp) mapping to transcriptional start sites (TSSs) and along GC-rich gene regulatory regions, and ones that are distributed along transcribed gene bodies downstream from the TSS with much longer median lengths of ~1.5 kb (Castillo-Guzman and Chédin 2021). Each have different properties and are preferentially detected by different methods (Castillo-Guzman and Chédin 2021). Experimental techniques to map structures throughout the lncRNA transcriptome have proved difficult owing to the low levels of lncRNA expression, limitations associated with the required sequencing depth, and large variability in results from different cell lines and different biological contexts owing to the high tissue and cell type specificity of lncRNAs.

In this study, we compare the frequency of computationally predicted G4 (pG4)-, iM (piM)-, and RL (pRL)-forming sequences across the lncRNA transcriptome with the protein-coding mRNA transcriptome using G4-iM Grinder and QmRLFS-finder. Both G4-iM Grinder and QmRLFS-finder are suitable for predicting RNA secondary structures and offer a wider range of detection for the structures of interest in comparison to many other RNA structure prediction tools. Despite the advances in availability and effectiveness of G4 prediction tools and the updates to the human reference transcriptomes, predictions of G4s in the lncRNA transcriptome have not been specifically addressed since 2012, and lncRNA G4s are, in general, less studied than their mRNA counterparts (Jayaraj et al. 2012). We additionally investigate the frequency of the predicted structures in different tissues using the lncRNA spatial atlas (lncSpA) "integration" lncRNA tissue expression data set and explore the potential functions of these tissue-associated lncRNA secondary structures (Lv et al. 2020). We then validate the formation of pG4 and piM structures for a select number of lncRNA sequences through a series of in vitro biophysical experiments.

Results

Predicted G4, iM, and RL formation in the human transcriptome

We first predicted the formation of G4, iM, and RL structures throughout the lncRNA transcriptome from the GENCODE v34 database using G4-iM Grinder and QmRLFS-finder algorithms. A total of 48,479 transcripts across 17,960 unique genes from the GENCODE lncRNA v34 database and 103,155 transcripts across 20,368 genes from the GENCODE mRNA v36 database were entered into the G4-iM Grinder and QmRLFS-finder prediction algorithms (Jenjaroenpun et al. 2015; Frankish et al. 2019; Belmonte-Reche and Morales 2020). For G4-iM Grinder results, a score threshold of 40 or more was set to define pG4s and a score threshold of −40 or less was set to define piMs as used previously (Belmonte-Reche and Morales 2020). No score parameter as a measure of the likelihood of RL formation in vitro or in vivo is provided by QmRLFS-finder, and thus, all pRL-forming motifs identified were included for subsequent analysis. In total, 6645 pG4-, 8032 piM-, and 6299 pRL-forming motifs were detected across 2964, 3388, and 1852 unique lncRNA genes, respectively, compared with 29,049 pG4s, 49,429 piMs, and 41,770 pRLs across 7912, 9182, and 7292 unique genes in mRNA (Table 1). Within both databases, most genes were found to contain fewer than 20 structures (Fig. 1A). Of note, we detected almost three times more pG4-forming sequences in lncRNA than the previous in silico lncRNA G4 predictions study, which detected 2394 total quadruplex-forming motifs using Quadfinder (Jayaraj et al. 2012). This increase is attributed to the properties of modern G4 prediction algorithms, such as Quadron, G4RNA Screener, and G4-iM Grinder, which detect non-canonical G4-forming sequences by allowing the presence of bulges and longer loops, and the increasing size of the annotated lncRNA transcriptome (Garant et al. 2017; Sahakyan et al. 2017a; Belmonte-Reche and Morales 2020). The even larger number of predicted iMs is surprising given that RNA iMs tend to be unstable (Lacroix et al. 1996). As a control, the G4-iM Grinder and QmRLFS-finder predictions were repeated after scrambling the sequence for

Table 1. Total number and percentage of predicted structures: pG4s, piMs, and pRLs from lncRNA and mRNA data sets

	Structure type	Total predicted structures	Unique genes	% of input genes	Unique transcripts	% of input transcripts	Unique structures
lncRNA	G4	6,645	2964	16.50	5,023	10.36	4,462
	iM	8,032	3388	18.86	6,077	12.54	5,101
	RL	6,299	1852	10.31	2,845	5.87	4,836
mRNA	G4	29,049	7912	38.82	20,046	19.43	14,054
	iM	49,429	9182	45.08	27,543	26.70	20,172
	RL	41,770	7292	35.80	17,649	17.11	24,622

each lncRNA. As expected, a reduction in the number of predicted structures was observed with 1286 pG4s, 2752 piMs, and 672 pRLs detected, confirming a strong dependence of the nucleotide sequence over other factors such as transcript length, structure density, or nucleotide composition.

To account for the notably higher number of transcripts/genes in the input data set for mRNA compared with the lncRNA data set, we calculated the percentage of unique transcripts and genes with predicted structure formation and observed a higher percentage of mRNA genes and transcripts containing predicted structures than lncRNA across all structure types (Table 1). The difference between lncRNA and mRNA was least pronounced in pG4s with a 2.35-fold increase, whereas the percentage of genes with piMs and pRLs was altered by 2.39- and 3.47-fold, respectively. Furthermore, the percentages of pG4- and pRL-containing transcripts are more similar within mRNAs, at 19.43% and 17.11%, respectively, compared with lncRNAs, at 10.36% and 5.87%. Next, we investigated the location of the predicted structural motifs in relation to the parent transcript (Fig. 1B). For lncRNA, a higher density of putative structure motifs was observed at the 5′-end of the transcripts, whereas mRNA transcripts contained peaks at both the 5′- and 3′-ends, representing the 5′ and 3′ untranslated

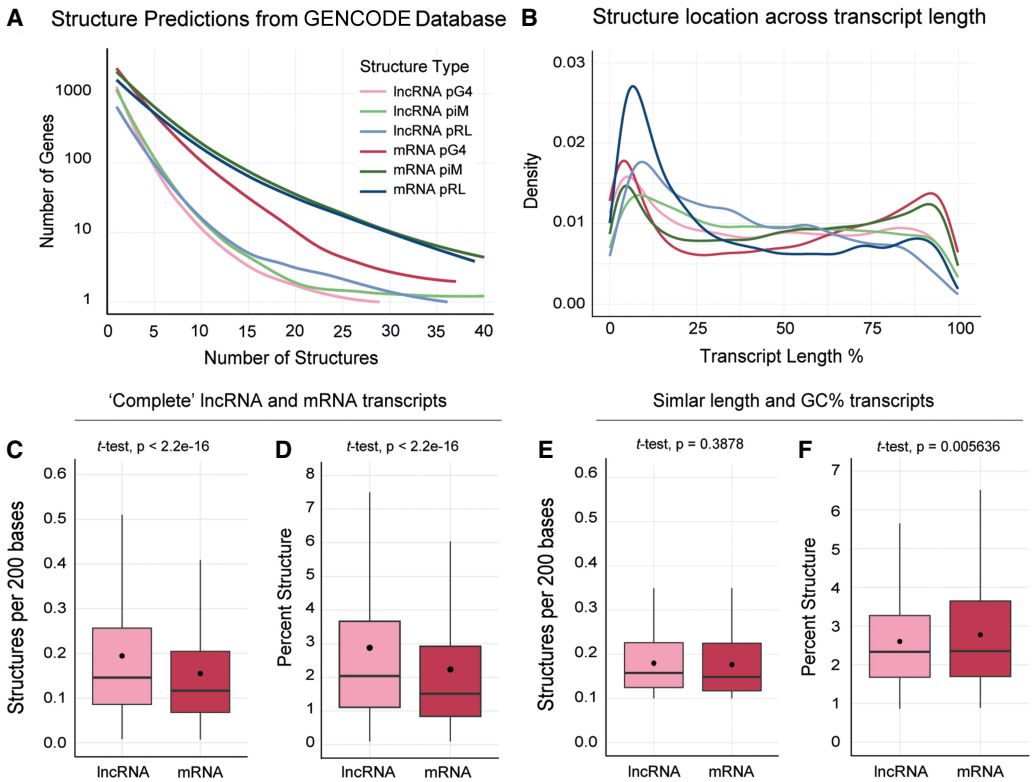


Figure 1. Comparisons between predicted structures in lncRNA and mRNA data sets. (*A*) Line graph showing the density of each structure in unique genes. The number of structures is limited to a maximum of 40 in a single gene for clarity (for full graph, see Supplemental Figure S1A). (*B*) Line graph showing the distribution of pG4, piM, and pRL structures across transcript length (5′ to 3′). Figure legend same as for panel *A*. (*C*,*D*) Boxplots showing the density of pG4s within the transcript (*C*) and the total percentage of the transcript consisting of pG4s (*D*) for the complete lncRNA and mRNA transcriptomes. Outliers removed for clarity, full graph in Supplemental Figure S1B,C. (*E*,*F*) Boxplots showing the density of pG4s within the transcript (*E*) and the total percentage of the transcript consisting of pG4s (*F*) for the lncRNAs and mRNAs with near-equivalent transcript length and GC content. For length standardization in *C* and *E*, the number of pG4/piM/pRL structures within a given transcript was divided by the transcript length to get the number of structures per base and then multiplied by 200. The number of bases used here is arbitrary and does not affect comparisons made between lncRNA and mRNA data sets.

regions (UTRs) of the transcripts. Comparing structure types, the incidence of piMs showed the most even distribution, whereas pG4s were preferentially localized at both the 5′- and 3′-ends for mRNA but only the 5′-end for lncRNA. G4s at the 5′ and 3′ UTRs of mRNAs are known to play regulatory roles in cap-dependent translation, post-transcriptional processing, and RNA localization (Lyu et al. 2021). pRLs showed the highest density at the 5′-end and lowest density at the 3′-end out of all the structures examined, in agreement with previous experimental studies mapping RLs near transcription start sites (Dumelie and Jaffrey 2017).

As the average length of mRNA transcripts is higher than that of lncRNA transcripts and given the large variation of lncRNA lengths, we standardized the number of each structure type predicted by the transcript length and percentage of nucleotides involved in the structure formation (Fig. 1C,D; Supplemental Fig. S1B,C). The number of structures per 200 nucleotides (nt) was used for length standardization as this is generally considered the minimum length for a lncRNA (Statello et al. 2021). Although there is a larger fraction of mRNAs containing one or more predicted G4s, iMs, and RLs, the lncRNA transcripts that did contain predicted structures had a higher density of pG4- and pRL-forming sequences (Table 1; Fig. 1C; Supplemental Fig. S1B). mRNA transcripts had a slightly higher density of piM-forming sequences. Next, we calculated the percentage of nucleotides of any given transcript involved in structure formation for each structure type and noticed a significantly higher proportion for RLs in both lncRNA and mRNA (Supplemental Fig. S1C). This result was expected as the median RL length of approximately 300 bases is significantly longer than the length of the G4 and iM structures, and RLs can also form from nascent RNA, therefore involving most of the newly generated transcript (Malig et al. 2020). The mean percentage of nucleotides throughout the transcript involved in structure formation was higher in lncRNA for all structure types investigated (Fig. 1D; Supplemental Fig. S1C).

Additionally, as inherent differences in the transcript length and sequence composition of the lncRNA and mRNA transcripts may impact the prediction results, we compared subsets of 14,637 lncRNA and 13,695 mRNA transcripts with near-equivalent transcript length (600–2000 nt) and GC content (41%–50%) (Supplemental Table S1). In comparison to the whole-transcriptome results above, the number and percentage of pG4- and piM-containing transcripts are notably more similar between lncRNAs and mRNAs (1.08× and 0.84× difference from lncRNA to mRNA, respectively) than pRLs, which remained approximately 3.4× less frequent in lncRNA than mRNA. A similar trend was observed when examining the number of predicted structures normalized to transcript length, with a similar density of pG4s and piMs between lncRNAs and mRNAs, whereas mRNAs had a higher density for pRLs (Fig. 1E; Supplemental Fig. S2A). The percentage of nucleotides in each transcript involved in structure formation was higher in mRNA than in lncRNA for all structure types examined (Fig. 1F; Supplemental Fig. S2B).

As any given RNA can potentially contain multiple different structures in different locations or compete with other structure types in overlapping sections, it is important to investigate these intersections to get a more comprehensive idea of the overall RNA structure. Furthermore, this concept is not accounted for in the total number or percentage of genes given in Table 1. We thus identified subsets of genes that contained only one structure type and subsets in which multiple types of structures were predicted to form in both the lncRNA and mRNA data sets. The distribution of the structure types showed marked differences between the lncRNA and mRNA data sets: 61.8% of lncRNAs contained only a single type of predicted structure in contrast to the mRNAs, which are more likely to be able to form multiple types of secondary structures, representative of the different functions of the structures (Supplemental Fig. S1D,E). Specifically, in lncRNA, of the 3338 genes containing at least one piM, 1291 also contained at least one pG4. Secondary structures in mRNA often tend to hinder translation, whereas in lncRNA, they act as topological marks recruiting proteins and other binding partners (Gaspar et al. 2013; Zampetaki et al. 2018; Mauger et al. 2019; Chillón and Marcia 2020; Graf and Kretz 2020; Varshney et al. 2021). The low numbers for the combination of pRL- and piM-forming sequences in the same gene were expected given the propensity of RLs to form within G-rich sequences as opposed to C-rich sequences.

As G-rich sequences are known to be essential for both G4 and RL formation and both G4-iM Grinder and QmRLFS-finder use this feature to predict structure formation, we additionally investigated overlaps in the sequences between pRLs, pG4s, and piMs in both the lncRNA and mRNA data sets (Supplemental Fig. S3). We found that 48.9% of the pRL-forming sequences had an overlap of at least 1 nt with pG4-forming sequences, but only 7.1% overlapped with piM-forming sequences. pRL-pG4 overlapping regions also involved more nucleotides, whereas a maximum of 4 nt overlapped for RLs and iMs. Similar results were seen for the mRNA data set. As a control, we also examined the frequency of overlap between pG4 and piM sequences and detected only nine instances representing 0.1% of each data set, with a maximum overlap of 1 nt, as expected given the requirement for highly G-rich or C-rich sequences to predict a G4 or iM, respectively.

The G4-iM Grinder algorithm also provides a score relating to the probability of the identified sequence to form the structure of interest; however and most importantly, its accuracy for RNA structures has not been experimentally validated. No such score parameter is currently included in the QmRLFS-finder algorithm for RLs. No significant differences were observed between the score distribution of pG4s and piMs between lncRNA and mRNA or between the different structure types, indicating a similar likelihood of formation (Supplemental Fig. S3A).

Structure-containing lncRNA tissue expression and pathway analysis

An interesting property of lncRNAs is their high tissue specificity relative to other forms of RNA (Cabili et al. 2011; Hon et al. 2017). We compared our G4-iM Grinder and QmRLFS-finder results with the lncSpA integration lncRNA tissue expression data set to examine if the presence of noncanonical structures correlated with a particular tissue type. Out of the 17,960 lncRNA genes from the GENCODE human v34 database, 7497 (41.7%) were present in the lncSpA tissue data set, of which 2152 were identified in our structure predictions data. All tissues with more than 10 associated lncRNAs contained numerous lncRNAs with at least one predicted noncanonical secondary structure, indicating potential roles of G4s, iMs, and/or RLs in lncRNA function across a diverse range of tissue and therefore cell types. The testes and brain were identified as tissues with the highest number of predicted structure-containing lncRNAs (Supplemental Fig. S3). Importantly, the testis and brain are tissues in which the majority of lncRNAs have been characterized, in part owing to their significant overexpression of lncRNAs relative to other tissues, and thus, the number of structure-containing lncRNAs was normalized to the total

number of lncRNA genes represented in each tissue from the lncSpA data set (Supplemental Fig. S5). Complex tissues such as the brain and pituitary contained high numbers of both lncRNA genes and predicted secondary structures (Supplemental Figs. S4, S5). Previous studies have identified ~40% of all lncRNAs are specifically expressed in the nervous system, and the involvement of lncRNAs in brain-related disorders, including pituitary adenoma, is increasingly evident (Derrien et al. 2012). Furthermore, many rG4s are known to be implicated in the development and progression of numerous neurodegenerative conditions (Simone et al. 2015; Wang et al. 2021). The secondary structures in lncRNAs identified in this paper represent potential biomarkers and therapeutic targets for these diseases. We also examined the likelihood of the pG4s and piMs to form in the given tissues by comparing the G4-iM Grinder Score parameter; no distinct pattern was observed, with all tissues showing similar means and distributions (Supplemental Fig. S6).

We explored the biochemical pathways associated with the lncRNA genes containing pG4, piM, or pRL secondary structures through Gene Ontology (GO) pathway analysis. Thirty-one of the 5459 genes have known functions, including the highly studied lncRNAs *NEAT1*, *MALAT1*, *XIST*, *H19*, and *MEG3*. Given the low rate at which lncRNAs are annotated and functionally characterized, these low counts are not surprising. Furthermore, 2152 of these lncRNA genes had tissue annotation, of which 11, mostly expressed in the brain followed by the testis and pituitary, have known cellular functions (Table 2). A small number of these lncRNAs that have been previously linked to particular biological pathways were also detected in the placenta, skin, colon, small intestine, and heart (Table 2). Secondary structures within these pathway-associated lncRNAs have previously been shown to coordinate lncRNA–protein interactions, highlighting their importance for function (Pintacuda et al. 2017; Uroda et al. 2019; Simko et al. 2020; Ghosh et al. 2022). Additionally, the formation of a DNA G4 in the *H19* gene has been shown to regulate transcription of the *H19* lncRNA with functional consequences (Fukuhara et al. 2017). Our prediction results indicate an rG4 also forms in the *H19* lncRNA and may have functional consequences; however, further in vitro and in vivo experiments are required to validate this. Our predictions identified a piM in the *CDKN2B-AS1* lncRNA. Previously, Ou et al. (2020) showed that *CDKN2B-AS1* can form an RNA–DNA triplex structure with the *CDKN2B* gene promoter instead of, or potentially in competition with, the predicted iM. Taken together, these data highlight the connection between secondary structure formation in lncRNA and its importance to the biological functions of lncRNA. As more lncRNAs are characterized, examining structure as a regulatory and functional mechanism becomes increasingly relevant.

Biophysical experiments confirm prevalent G4 formation in lncRNA

For biophysical testing of lncRNA pG4s, we relaxed the G4-iM Grinder score threshold to 20 or more and MinNRuns to three to include nontraditional G4s with mismatches, bulges, and alternative loops, which may present with a lower score. This lower threshold has previously been used to identify pG4s in SARS-CoV-2 and other viral genomes with "medium" probability of in vitro formation and thus explore the correlation between score and structure stability (Belmonte-Reche et al. 2021). An additional 75,639 pG4s were detected across 23,278 transcripts in 8300 unique genes. To validate structure formation from the predictions, we ran various in vitro biophysical analyses on 10 pG4 and seven piM lncRNA sequences across different score brackets, nucleotide compositions, G/C runs, and lengths. Candidate pG4 sequences were selected based on a mix of their high prevalence throughout the transcriptome and distribution across different score brackets. As the formation of RNA iMs is contested, candidate piMs were selected based on motifs with the highest probability of in vitro formation determined by the most negative scores that were not poly(C) repeat sequences. The minimal rG4/iM motif was used without any flanking nucleotides. pRL sequences were

Table 2. Tissue expression and Gene Ontology (GO) pathway analysis of selected lncRNA genes containing pG4, piM, or pRL structures

Gene symbol	Ensembl Gene ID	Predicted structure type(s)	Tissue(s)	Main pathway(s) identified
MIAT	ENSG00000225783	G4, iM, RL	Brain, pituitary	Cell fate specification/commitment, differentiation
H19	ENSG00000130600	G4, iM, RL	Placenta	Metabolism of purines, oxidative stress, ATP metabolism, hepatocyte apoptosis
ELFN2 (lncRNA variant)	ENSG00000243902	G4, iM, RL	Brain, testis	Regulation of protein dephosphorylation
HOTAIR	ENSG00000228630	G4	Testis, skin	Epigenetic silencing of gene expression
MYCNOS	ENSG00000233718	G4, iM	Testis	Negative regulation of protein phosphorylation
LINC01783	ENSG00000233421	G4, RL	Testis	Molecular adaptor activity—targets miRNA to promote cancer-associated pathways
LINC03126	ENSG00000228549	G4, RL	Testis	Molecular adaptor activity
SNAP25-AS1	ENSG00000227906	iM	Brain, pituitary	SRP-dependent cotranslational protein targeting to membrane (endoplasmic reticulum), signal sequence recognition
CDKN2B-AS1	ENSG00000240498	iM	Colon, small intestine	Epigenetic silencing of gene expression
GJD2-DT	ENSG00000250007	iM, RL	Brain, heart, pituitary	SRP-dependent cotranslational protein targeting to membrane (endoplasmic reticulum), signal sequence recognition
TAB2 (lncRNA variant)	ENSG00000228408	RL	Brain	MYD88-dependent TLR signaling pathway, JNK cascade, response to IL1 regulation of I kappa B kinase/NF-kB signaling, MAPK signaling pathway

not tested owing to experimental constraints regarding the RNA–DNA hybrid nature of the structure and difficulties associated with the length of the sequences. Details of all sequences tested are given in Table 3.

We initially investigated the formation of G4 structures in the lncRNA pG4 oligos alongside three positive control G4 sequences (*TRF2*, *ALS-FTD*, *TERRA*) and a mutated negative control non-G4 sequence (*TRF2* mut) using Thioflavin T (ThT) and *N*-methyl mesoporphyrin IX (NMM) fluorescence assays. Both ThT and NMM are known to increase in fluorescence upon binding to a quadruplex (Xu et al. 2016; McBrayer et al. 2019). All tested sequences showed significantly increased fluorescence over the negative control ($P<0.0001$) and had similar fluorescence signals to the positive control sequences, strongly suggesting G4 formation in all tested lncRNA sequences (Fig. 2A).

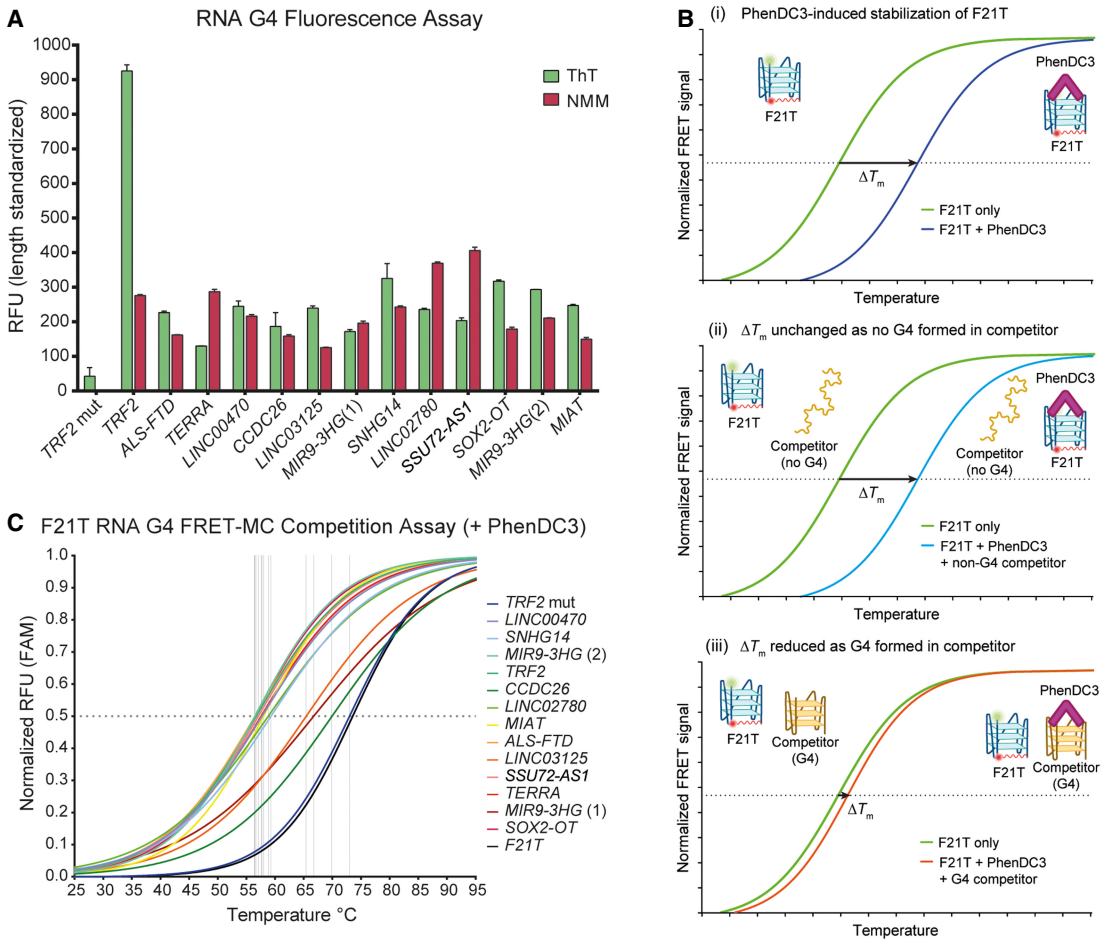
To corroborate these results, we performed a fluorescence resonance energy transfer melting competition (FRET-MC) assay to detect G4 formation (Luo et al. 2021). The competition assay measures the ΔT_m of a fluorescent control G4, F21T, in the presence of excess test lncRNA oligo upon addition of PhenDC3 G4-stabilizing ligand (Fig. 2B). If a G4 forms in the test oligo, PhenDC3 binds to the excess test oligonucleotides, leading to a decrease in the level of F21T G4 stabilization and thus ΔT_m, expressed as a low (close to zero) S-factor. We first confirmed the T_m of F21T in the presence of all competitors without PhenDC3 was equal to the F21T alone as a baseline, confirming that the addition of the competitor itself does not impact the T_m of F21T (Supplemental Fig. S7). With PhenDC3, all lncRNA oligos tested except *CCDC26*, *MIR9-3HG* (1), and *LINC03125* showed G4 formation with a decrease in the F21T ΔT_m, similar to the positive controls *TRF2*, *ALS-FTD*, and *TERRA* (Fig. 2C). For the *TRF2* mut negative control, no change in the F21T ΔT_m was detected (Fig. 2C). *CCDC26*, *MIR9-3HG* (1), and *LINC03125* resulted in smaller decreases in T_m, with the melting curves in between the positive and negative controls (Fig. 2C, D). Mirroring the melting curves, most of the lncRNA pG4 oligos had an S-factor value close to zero, indicating G4 structure formation, whereas the *TRF2* mut negative control had an S-factor value of 0.91, close to the F21T alone control (1.00) (Fig. 2D,E; Luo et al. 2021). *CCDC26* had a negative result in this experiment, and *MIR9-3HG* (1) and *LINC03125* lncRNA pG4 sequences showed inconclusive results with S-factors of 0.72, 0.55, and 0.48, respectively (Fig. 2D,E). In combination with the ThT and NMM fluorescence results for these lncRNA sequences, we suggest that these G4s may have lower stability or exist in competition with an alternate secondary structure such as a hairpin. It is also possible that the resulting G4s may have reduced affinity for PhenDC3. We consider this hypothesis unlikely as two other ligands, NMM and ThT, bind reasonably well to these structures.

To further validate and characterize G4 formation, we performed isothermal difference spectra (IDS), thermal difference spectra (TDS), UV-melt, and circular dichroism (CD) (Fig. 3). The lncRNA sequences *SSU72-AS1*, *LINC02780*, *MIR9-3HG* (2), *SOX2-OT*, *MIR9-3HG* (1), and *LINC00470* all displayed the characteristic G4 peaks in IDS and TDS at ~245 and ~275 nm and the trough at ~295 nm (Fig. 3A,B; Mergny et al. 2005; Timmer et al. 2014). IDS for *SNHG14* and *MIAT* resulted in negative Δ absorbance values

Table 3. Sequences of oligonucleotides used for biophysical studies

RNA	Sequence (5′ to 3′)	No. of bases	No. of G/C runs	Score (Grinder/G4H)[a]
SNHG14	GAGGCUUUGGGGUCUGGGCAGGG	23	4	46/1.56
LINC02780	GGGAGGCUGAGGUGGG	16	3	44/1.64
SSU72-AS1	GGGCGUGGUGGCGGG	15	3	45/1.68
SOX2-OT	GGGAGAGGGAAUGGGAGGGUAGAAGAAGGGGG	32	5	54/1.84
MIAT	GGGCCUGGGGAGGGGGCCUGGG	22	4	61/2.08
LINC00470	GGUGGUGGAGGGAGCCCUGGGUCGAGUGG	29	5	30/0.92
CCDC26	GGGUUCCUAGGUGAUAGGGUGUGG	24	4	32/1.00
LINC03125	GGAGGCUCUUCUUUAGGGGAGAGAGGGAAAUGGG	34	4	38/1.20
MIR9-3HG (1)	GGGACUUGGCAGGGGACCUGGG	22	3	45/1.44
MIR9-3HG (2)	GGGGGAUGGGGAACAGGGCUGUGAAUGGG	29	4	56/1.88
ALS-FTD	GGGGCCGGGGCCGGGGCCGGGG	22	4	67/2.36
TERRA	AGGGUUAGGGUUAGGGUUAGGG	22	4	50/1.64
TRF2	CGGGAGGGCGGGGAGGGC	18	4	67/2.22
TRF2 mut	CGUGAGUGCGCUGAGGGC	18	1	N/A/0.61
F21T	FAM-GGGTTAGGGTTAGGGTTAGGG-TAMRA	21	4	50/1.71
CYP4F26P	CCCUCCCUCCCAGCCCCAUGAUUACCCC	28	5	−58/−2.04
LY86-AS1	CCCCAAACUGCCCCUGCUCCCCUCUCCCCC	30	4	−64/−2.28
TMEM72-AS1	CCCCACAGGCCCUCCCCUCCCC	22	4	−66/−2.44
SLC12A5-AS1	CCCCCAACCCCCUCCGCCCGCCCGCCCCC	29	5	−73/−2.72
LNC-LBCS	CACCAUCCCCCCCACCCCCCCACCCCACCC	30	5	−66/−2.88
GATA6-AS1	CCCCGACCCCACCCCCUACCCCCCGCCC	28	5	−76/−2.96
MYO3B-AS1	CCCGCCCCGCCCCGCCCCCCACCCCC	26	5	−80/−3.16
Tel21	CCCTTACCCTTACCCTTACCC	21	4	−50/−1.71

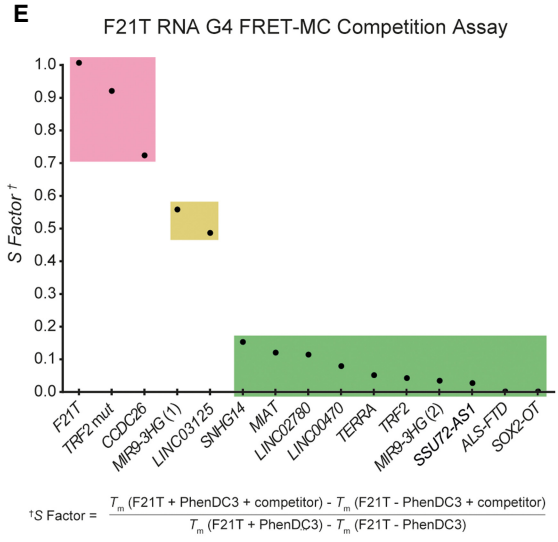
[a]Prediction tools: (Grinder) G4-iMGrinder, (G4H) G4 Hunter.



D

	Oligo	ΔTm of F21T*	S-factor	Score**
F21T without competitor		18.53	1.00	50
F21T + Competitor Oligo of Interest	*TRF2* mut	16.95	0.91	-
	CCDC26	13.31	0.72	32
	MIR9-3HG (1)	10.26	0.55	45
	LINC03125	8.94	0.48	38
	SNHG14	2.79	0.15	46
	MIAT	2.19	0.12	61
	LINC02780	2.07	0.11	44
	LINC00470	1.43	0.08	30
	TERRA	0.92	0.05	50
	TRF2	0.75	0.04	67
	MIR9-3HG (2)	0.60	0.03	56
	SSU72-AS1	0.47	0.03	45
	ALS-FTD	-0.13	-0.01	67
	SOX2-OT	-0.20	-0.01	54

*ΔT_m of F21T = T_m (F21T + PhenDC3 + competitor) - T_m (F21T - PhenDC3 + competitor)
**Score = G4-iM Grinder Score (combination of G4Hunter and PQSFinder)



†S Factor = $\frac{T_m \text{ (F21T + PhenDC3 + competitor)} - T_m \text{ (F21T - PhenDC3 + competitor)}}{T_m \text{ (F21T + PhenDC3)} - T_m \text{ (F21T - PhenDC3)}}$

Figure 2. Fluorescence-based assays detecting G4 formation in lncRNA pG4 oligos. (*A*) ThT and NMM fluorescence normalized to sequence length. (*B*) Schematic of G4 FRET-MC experiment. (*i*) Initially, the F21T fluorescent G4 and nonfluorescent pG4 oligo (competitor) undergo a melt curve in the absence of the G4-stabilizing ligand, PhenDC3, and the T_m of the F21T G4 is determined. Next, the melt curve is performed in the presence of PhenDC3 to determine the ligand-induced stabilization of F21T, ΔT_m. (*ii*) In the event that the competitor does not form a G4, the PhenDC3 ligand stabilizes the F21T G4, and the ΔT_m remains unchanged. (*iii*) In the event of a G4 forming in the competitor, the PhenDC3 ligand stabilizes the competitor, and the ΔT_m of the F21T is reduced. In all cases, only the T_m of the F21T fluorescent G4 is measured. (*C*) Nonlinear fit of normalized F21T melt curves ± competitor oligos with PhenDC3. Horizontal dotted line at $y = 0.5$, vertical dotted lines at T_m of F21T in each sample. (*D*) Table showing ΔT_m of F21T upon addition of PhenDC3 ± competitors, S-factor, and G4-iM Grinder score of samples. (*E*) S-factor analysis showing the relative PhenDC3-induced stabilization of F21T ± competitors. Samples with results positive for G4 formation highlighted in green, inconclusive or intermediate G4 formation highlighted in yellow, and negative for G4 formation highlighted in red.

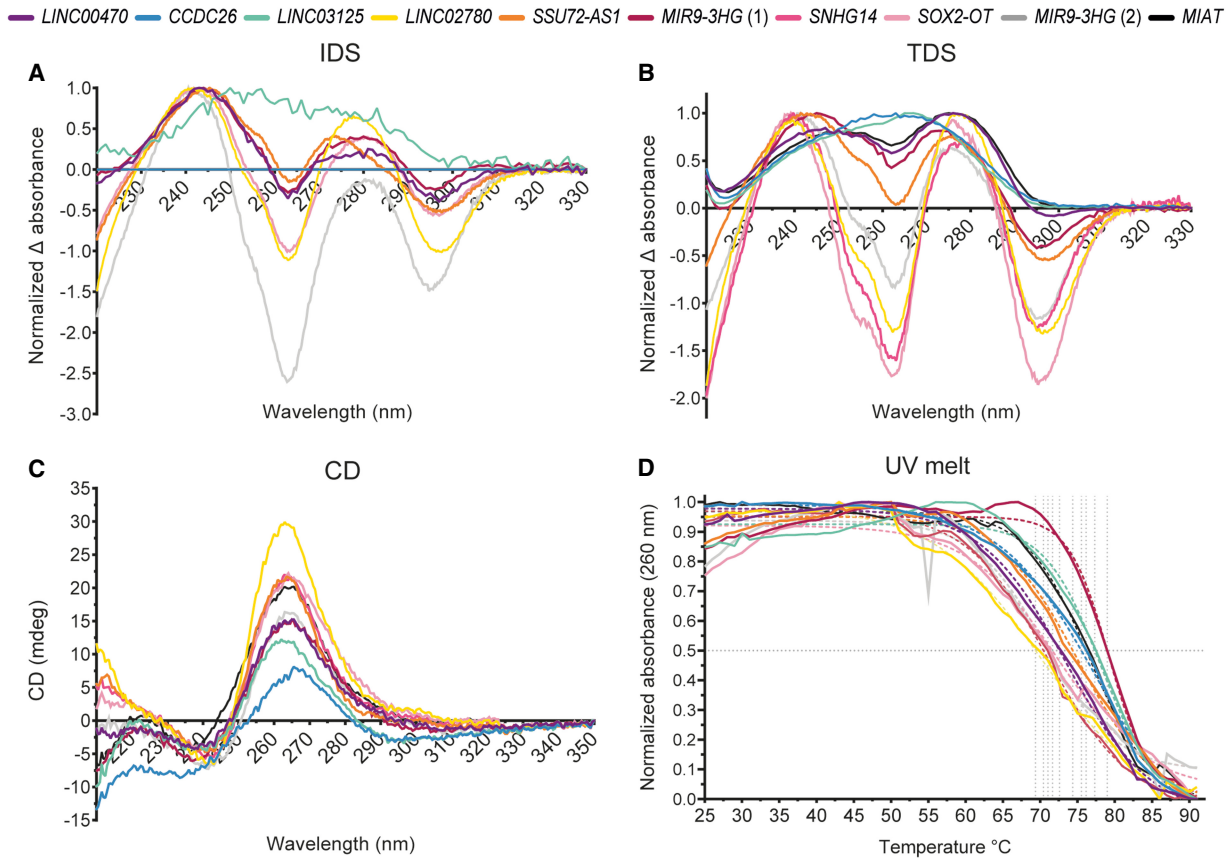


Figure 3. pG4 lncRNA oligos show characteristic features of rG4s. (*A*) IDS showing normalized ΔA from 220 to 330 nm between unfolded (no K^+) and folded (10 mM K^+) conditions at 37°C. (*B*) TDS showing normalized ΔA from 220 to 330 nm between unfolded (95°C) and folded (37°C) in 10 mM K^+ buffer. (*C*) CD spectra from 210 to 350 nm for G4 topology. (*D*) UV melt showing normalized absorbance at 260 nm from 90°C to 25°C, cooled at a rate of 1°C/min. Colored dashed lines show nonlinear fit of normalized data. Individual melt profiles (raw data) and T_m for each oligo tested are given in Supplemental Figure S10.

for all wavelengths measured and could not be properly normalized; raw data are provided in Supplemental Figure S8. Because of the high stability of rG4s, it is possible that a quadruplex structure in these sequences was forming in the "unfolded" condition without K^+ and thus skewed the shift in absorbance. We determined the topology of the G4 structures using CD (Fig. 3C; Kypr et al. 2009). All sequences displayed a CD spectra characteristic for parallel quadruplexes, indicated by a negative peak at ~245 nm and a dominant positive peak at ~265 nm, as is expected in RNA owing to the increased steric hinderance from the ribose backbone (Fig. 3C; Paramasivan et al. 2007; Kypr et al. 2009). Finally, we examined the thermal stability of the G4s with a UV-melt assay at 260 nm (Fig. 3D). As expected for rG4s, all the tested sequences showed high stability, with the lowest T_m at 69.0°C and highest T_m at 78.9°C in the presence 10 mM K^+ (Fig. 3C; Supplemental Fig. S9). In cells the concentration of K^+ is approximately ten times higher, which would further stabilize the structures, increasing the likelihood of their formation in vivo and suggesting that the G4s are important to the lncRNA functions. The *CCDC26* and *LINC03125* sequences had some of the highest T_m values, suggesting that the high S-factor values for these lncRNA sequences from the FRET-MC competition assay are more likely owing to reduced affinity to PhenDC3 rather than low stability. To detect the formation of intermolecular G4 formation in the samples, the lncRNA pG4 sequences were resolved with 20% native polyacrylamide gel electrophoresis, stained with ThT and NMM to visualize G4s, and counterstained with SYBR safe gel stain for total nucleic acids (Supplemental Fig. S10). Multiple bands were seen for lanes corresponding to *LINC00470*, *MIR9-3HG* (2), and *SNHG14*, which indicate the formation of multiple structures or multimeric G4s in the sample.

Overall, eight of the 10 lncRNA pG4 oligos tested formed a G4 structure evidenced by the ThT and NMM fluorescence, FRET-MC competition assay, TDS, IDS, and CD regardless of the differences in score, lengths, and nucleotide composition of the sequences, indicating that the majority of the pG4s identified in the lncRNA prediction data set are highly likely to form. Two of the sequences selected are present in multiple lncRNAs, suggesting that they may have a specific role in contributing toward lncRNA interactions with other molecules such as protein binding partners. Specifically, the sequence selected for *LINC02780* is also present in 104 other lncRNAs, and the sequence selected for *SSU72-AS1* is present in 66 other lncRNAs (Supplemental Table S2). Given the enrichment of lncRNAs in repeat elements and transposable elements (TEs) of the genome, we further investigated the prevalence of the computationally predicted lncRNA pG4 motifs within these regions using the RepeatMasker database retrieved from the UCSC Genome Browser (Kent et al. 2002; Smit et al. 2015; Mattick et al. 2023). In the top 10 most common pG4 sequences (with G4-iM Grinder parameters set to score 20 or more;

MinNRuns, three or more), up to 56.7% of the occurrences were harbored within repeat elements (Supplemental Table S3). In comparison, an average of 20.1% of all the lncRNA pG4s identified with the same G4-iM Grinder parameters were contained within repeats. Additionally, under high-confidence G4-iM Grinder parameters (score 40 or more; MinNRuns, four or more), an average of 34.2% of lncRNA pG4s were harbored in repeat elements compared with 25.8% for mRNA pG4s. These results may indicate a role for rG4 structures in functions of lncRNAs that are mediated by highly evolutionarily conserved repeat domains by facilitating lncRNA interactions with proteins.

Z-RNA formation in C-rich piM sequences

Our G4-iM prediction results identified a very high frequency of potential iM-forming sequences. This is surprising because RNA iMs have previously been reported to have very poor stability in vitro and their in vivo formation is highly debated (Lacroix et al. 1996; Snoussi et al. 2001). To account for this, we selected lncRNA sequences with highly negative G4-iM Grinder scores (−58 to −80) for biophysical studies to maximize their probability of formation. In contrast to the G4 results, none of the C-rich, piM RNA oligos displayed the characteristic TDS or CD peaks expected for iMs under any of the pHs tested, indicating that even these strong-scoring candidates do not form stable iM structures (Fig. 4; Supplemental Fig. S11). These results agree with the existing literature reporting the poor thermal stability of RNA iMs, attributed to the steric hindrance caused by juxtaposed 2′-OH groups of the ribose sugars in the narrow grooves of the RNA (Collin and Gehring 1998). Our observations indicate that G4-iM Grinder grossly overestimates the ability of RNA sequences to form iM.

The TDS spectra at pH 5.5 and pH 6.0 are consistent with a Z-form nucleic acid structure with characteristic peaks at ∼245 and ∼275 nm and with the trough at ∼295 nm (Fig. 4A,B; Mergny et al. 2005). To form this left-handed double helix, the transcript may fold in on itself or form intermolecular structures (Herbert 2019). Increasing the pH to 7.0 destabilized the Z-RNA structure as shown by the loss of the 295-nm trough (Fig. 4C), indicating a role of pH-dependent cytosine protonation in Z-RNA formation. IDS, which detects structure formation upon addition of stabilizing cations, specifically K^+ for G4s, was not performed for these samples as cations do not specifically stabilize or induce formation of iMs and Z-RNA, with pH and other environmental factors being more important for iMs, and Z-RNA structures are known to be stabilized by cytosine methylation or other modifications that alter the conformation of *N*-glycosidic bonds (D'Ascenzo et al. 2016; Abou Assi et al. 2018; Guédin et al. 2018). However, IDS and TDS provide the same information on G4 folding. CD spectra for the piM samples are similar to previously reported experimental CD spectra for Z-RNA, with positive values at ∼260 to ∼280 nm and negative values at ∼230 to 240 nm (Supplemental Fig. S11; Miyahara et al. 2016). Z-RNA has mostly been studied in GC-rich sequences, particularly poly-GC repeats; however, their formation in C-rich sequences such as those tested here is significantly less studied. A few studies have shown Z formation in nucleic acids with non-poly-GC sequences, including Z-DNA formation in a nonalternating sequence with chemically modified cytosines (Wang et al. 1985). More recently, Z-like conformations have been found in many published RNA crystal structures with varying sequences, including CpC in which the 3′-cytosine is found in *syn*- and *anti*-conformation in equal measure, indicating the potential for C-rich sequences to exist in the alternating *syn–anti* pattern characteristic for Z-RNA (D'Ascenzo et al. 2016). Duplexes of C-rich RNA have been previously reported to form over iMs owing to greater stability (Collin and Gehring 1998). Additional hydrogen bonds from the ribose sugar 2′-OH and cytidine O2 interacting with water molecules increase the stability of the structure (Placido et al. 2007). In addition, the protonated cytosines at the lower pH conditions tested may be able to offset the negative charge of the RNA backbone, allowing some of the cytosine bases to be in *syn*-confirmation.

Discussion

Several potential G4, iM, and RL motifs were found in lncRNA using structure prediction algorithms. The density of these unusual structures was higher than in mRNA, suggesting potential roles of these motifs in noncoding RNAs. At a transcriptome-wide scale, G4 and RL structures may play a greater role in lncRNA function over iMs, which did not form in vitro, and may be more functionally relevant than in mRNA as they are predicted to occur at a higher density in lncRNA. However, further experimental work is required given the presence and in vitro formation of structural motifs does not necessarily attribute in vivo formation or functional significance within cells. More piM structures were detected than pG4s, despite the proportion of guanines and cytosines being ∼25% each as expected in both the lncRNA and mRNA transcriptomes. This may be because of a somewhat looser sequence requirement for piM identification compared with pG4s; however,

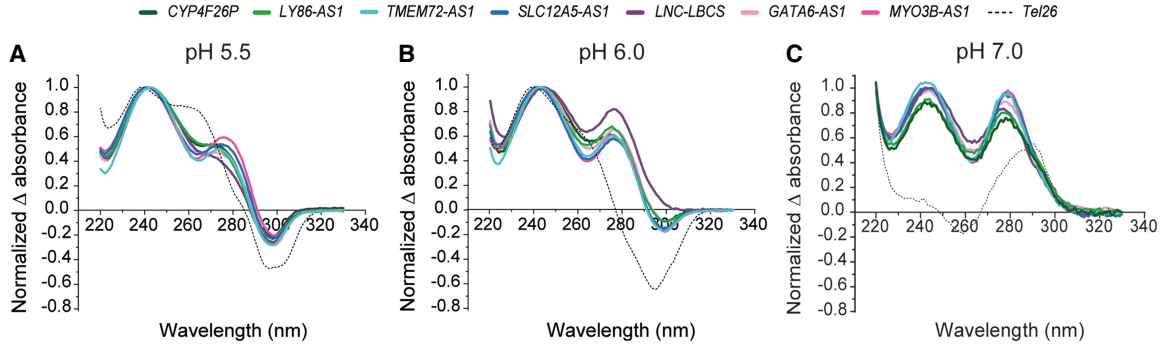


Figure 4. TDS of C-rich (predicted iM) lncRNA oligos show pH-dependent Z-RNA secondary structure formation. TDS showing normalized difference in absorbance from 220–330 nm between unfolded (90°C) and folded (20°C) conditions in 10 mM sodium cacodylate buffer at pH 5.5 (*A*), pH 6.0 (*B*), and pH 7.0 (*C*). *Tel26* (dashed line) shows positive control DNA iM.

we stress that these predictions do not account for many local environmental factors that have a greater effect on iM formation than G4 formation, such as pH (Irving et al. 2022). Thus, although a greater number of piMs than pG4s were identified, the number of actual iMs formed under physiological conditions is likely to be less, especially after factoring in results from the biophysical experiments in this study.

Nonetheless, G4-iM Grinder predictions were found to be fairly accurate for G4s as validated by our biophysical studies. The majority of lncRNAs tested here have not been previously reported to contain G4s. For example, in this paper, we show for the first time the formation of a highly stable G4 structure within the *LINC00470* and *SOX2-OT* lncRNAs in vitro, which may play an important role in their functions. In the past five years, *LINC00470* has been identified to coordinate glioblastoma cell autophagy, malignancy in gastric cancer cells, endometrial cancer progression, and proliferation and metastasis in melanoma, highlighting the biological relevance of this lncRNA in various types of cancer (Liu et al. 2018a,b; Yan et al. 2020; Huang et al. 2021; Yi et al. 2021). *SOX2-OT* lncRNA regulates the expression of a key transcription factor and oncogene, *SOX2*, a critical factor in maintaining pluripotency and embryonic and neural stem cells (Schaefer and Lengerke 2020; Mirzaei et al. 2022). Consequently, *SOX2-OT* participates in tumorigenic pathways for multiple cancer types, including lung cancer, pancreatic ductal adenocarcinoma, glioblastoma, osteosarcoma, triple-negative breast cancer, and more (Hou et al. 2014; Shahryari et al. 2015; Shafiee et al. 2016; Su et al. 2017; Zhang et al. 2017; 2022; Teng et al. 2019; Zhan et al. 2020; Herrera-Solorio et al. 2021; Liu et al. 2021; Mirzaei et al. 2022). Additionally, *SOX2-OT* plays an important role in vertebrate development and is inversely correlated with *SOX2* expression in embryonic stem cells during neurogenesis and differentiation of neural cells (Knauss et al. 2018; Messemaker et al. 2018). rG4s are similarly associated with cancers and stem cell differentiation and development, suggesting that the formation of the structure is likely to play a role in the disease association of *LINC00470*, *SOX2-OT*, and other lncRNAs (Varshney et al. 2020; Tateishi-Karimata and Sugimoto 2021; Ghafouri-Fard et al. 2022; Lyu et al. 2022; Zyner et al. 2022).

We identified several pG4 motifs that occur multiple times across different lncRNAs, which may permit multimerization to form higher-order structures and facilitate lncRNA–lncRNA interactions through the formation of intermolecular G4s (Frasson et al. 2022). Notably, intermolecular G4s can be sensitive to local RNA concentration, which may be limited to regions of high local transcription for lncRNAs that are often lowly expressed in vivo (Jain and Vale 2017). Predicting and validating the formation of intermolecular G4s remains challenging, with recent advancements from G4RP-seq (Yang et al. 2022). Additionally, highly prevalent pG4 sequences were found to be strongly associated with repeat regions and transposable in the human genome. Many previous studies have revealed the regulatory role of genomic repeats in lncRNA transcription, and repetitive elements within multiple lncRNAs have been shown to play important roles in their functions (Kapusta et al. 2013; Johnson and Guigó 2014; Fort et al. 2021; Mattick et al. 2023). Notably, G4 structures have previously been reported in TEs of the human genome and are particularly prevalent in evolutionarily young short interspersed nuclear elements (SINEs) and long interspersed nuclear elements (LINEs), in which they have been proposed to play a role in regulating transcription and genome evolution (Lexa et al. 2014; Sahakyan et al. 2017b; Makova and Weissensteiner 2023). In vertebrate RNAs, evolutionarily conserved structures were found to overlap with SINE, LINE, and simple repeats; however, investigations into specific secondary structures within these regions and their functional roles remain unclear (Seemann et al. 2017).

In contrast to the pG4s, the G4-iM Grinder predictions were not accurate for iMs, likely a consequence of the iM search parameters initially designed for DNA and not for RNA. Extreme caution is therefore needed when using G4-iM Grinder for RNA iMs. This software predicted a higher number of iMs than G4s in both mRNA and lncRNA, but even the highest-scoring iM candidates failed to form a stable iM structure under favorable conditions in vitro. Additionally, although the similar score distribution of pG4s and piMs indicated a similar likelihood of formation according to G4-iM Grinder, the large difference in the stability of the different structures is not accounted for in this parameter. rG4s are known to have greater thermal stability to their DNA counterparts, whereas the opposite is true for iMs (Collin and Gehring 1998; Saccà et al. 2005; Zaccaria and Fonseca Guerra 2018). Technologies assessing the cellular functions of iM structures have only just started to emerge, and it is possible that the development of new techniques combined with tailoring pre-existing experiments (e.g., ChIP-seq, RIP-seq, CLIP-seq) to target iMs will reveal more information on the potential existence—or inexistence—of these structures in RNA.

The piM lncRNA sequences instead formed a Z-RNA structure, as shown by characteristic Z-RNA features in TDS spectra. Z-RNA formation in these C-rich sequences was unexpected but is supported by the greater stability of C-rich RNA duplexes over RNA iMs, as well as a prior study in DNA in which Z-DNA is shown to form in chemically modified or non-GC repeat sequences (Wang et al. 1985; Collin and Gehring 1998). In cells, Z-RNA formation in these sequences may be different as a variety of physiological factors such as long-range RNA–RNA interactions and Z-RNA-binding proteins come into play. Given that Z-RNA interactions with proteins, specifically ADAR, have been shown to be important in RNA editing and human diseases, identifying these structures in alternative RNA types and sequences warrants further investigation. Although there are currently no predictive algorithms for Z-RNA, Z-hunt and DeepZ predict Z-DNA formation using different approaches (Ho et al. 1986; Beknazarov et al. 2020). Z-hunt evaluates sequences based on the energetic parameters of neighboring dinucleotides for B-DNA to Z-DNA transition, whereas DeepZ uses a deep learning approach by analyzing both Z-hunt predictions with results from various experimental studies, including ChIP-seq, epigenomics, and chromatin organization (Ho et al. 1986; Beknazarov et al. 2020). DeepZ predictions of Z-DNA-forming regions in the human genome identified introns as the largest subset followed by 5′ UTRs and promoters, indicating a role of these structures in regulating gene expression and potential Z-RNA formation in intronic noncoding RNAs (Beknazarov et al. 2020). Further work to assess Z-RNA formation from these intronic regions and the stability of the Z-RNA sequences identified in this study will provide better understanding of Z-RNA formation in the noncoding transcriptome.

In summary, we provide a comprehensive data set of predicted noncanonical secondary structures throughout the long noncoding transcriptome and connect these results to potential lncRNA functions. Complementing the predictions data with biophysical studies, we showed that G4-iM Grinder predictions were generally correct for RNA quadruplexes, and we validated formation of G4s in many lncRNAs for the first time. In contrast, we detected the formation of Z-RNA in C-rich sequences wrongly predicted to form iM by G4-iM Grinder.

Methods

Secondary structure predictions

lncRNA and mRNA transcript sequences and metadata were downloaded from the GENCODE human database v34 and v36, respectively (Frankish et al. 2019). We implemented the G4-iM Grinder algorithm to detect pG4- and piM-forming sequences throughout the transcriptome using RNA parameters: DNA = F, Complementary = F. The score threshold was set to 40 or more for G4 structures and −40 or less for iMs. A more positive score indicates a higher likelihood for G4 formation, and a more negative score indicates a higher likelihood for iM formation (Belmonte-Reche and Morales 2020). pRL-forming regions were identified using QmRLFS-finder (Jenjaroenpun et al. 2015). All downstream analyses were conducted in RStudio (R Core Team 2023). lncRNA tissue expression data were downloaded from the lncSpA v2.0 integration data set (Lv et al. 2020). Pathway analysis was performed using the clusterProfiler R package (Wu et al. 2021). Genomic repeat regions from RepeatMasker were retrieved from the UCSC Genome Browser (Kent et al. 2002; Smit et al. 2015).

Oligonucleotide preparation

Ten pG4 and seven piM sequences were selected across different score brackets for in vitro biophysical testing to confirm the structure formation. Synthetic RNA oligonucleotides mimicking these sequences were purchased from Eurogentec (Belgium) at 40-nmol scale with RP-HPLC purification and stock solutions prepared at 100 µM in nuclease-free water. The minimal G4/iM motif was used, without flanking nucleotides or modifications. The sequences of tested pG4 and piM motifs and all positive and negative controls used are provided in Table 3.

G4 fluorescence assays

ThT and NMM fluorescence assays were performed in 96-well plates using a M1000 Pro (Tecan) plate reader (Xu et al. 2016). G4 oligos were diluted to 7.5 µM in K100 buffer (100 mM KCl, 10 mM lithium cacodylate at pH 7.2) and annealed by heating for 5 min at 95°C, before being cooled to room temperature (~25°C) over 2 h. G4-forming TRF2, ALS-FTD, and TERRA oligos were used as positive controls, and TRF2 mut was used as a negative control. For each sample, 3 µM oligo was incubated with 2 µM of the fluorescent ligand (either ThT or NMM) in K100 buffer for 10 min at 25°C. Each sample was tested in duplicate with a final volume of 100 µL per well. Fluorescence emission was measured at λ_{ex} = 420 nm, λ_{em} = 490 nm for ThT and λ_{ex} = 380 nm, λ_{em} = 610 nm for NMM.

FRET-MC assay

FRET-MC experiments were performed in 96-well plates using a CFX96 RT-PCR instrument (Bio-Rad), as previously described (Luo et al. 2021). pG4 oligos were diluted to 7.5 µM in K10 buffer (10 mM KCl, 90 mM LiCl, 10 mM lithium cacodylate at pH 7.2) and annealed by heating for 5 min at 95°C and then cooling to room temperature over 2 h. The G4-forming TRF2, ALS-FTD, and TERRA oligos were used as positive controls and TRF2 mut as a negative control. For each sample, 3 µM of oligo was incubated with 0.2 µM of carboxyfluorescein (FAM)-tagged F21T human telomeric G4 oligo in the presence and absence of 0.4 µM of the G4-stabilizing ligand PhenDC3 (De Cian et al. 2007). The F21T sequence is provided in Table 3. Samples were incubated for 5 min at 25°C before measuring the fluorescence of F21T FAM-tag at 0.5°C intervals as the samples are heated to 95°C at 0.5°C/min. Formation of a G4 in the test oligo was confirmed by a decreased T_m of F21T in the presence of PhenDC3 owing to the excess test oligo pG4 outcompeting the binding between PhenDC3 and the fluorescent F21T G4, leading to decreased stabilization of the F21T G4. Each experimental condition was tested in duplicate with a final volume of 25 µL in K10 buffer. A schematic of the experiment is provided in Figure 2B.

IDS, TDS, and UV-melt

pG4 oligos were diluted to 3 µM in 10 mM lithium cacodylate buffer (pH 7.0) and heated for 3 min to 95°C before cooling to room temperature over 2 h. piM oligos were diluted to 3 µM in 10 mM sodium cacodylate (pH 5.5, 6.0, or 7.0) and similarly heated and cooled to anneal structures All measurements for IDS, TDS, and UV-melt were performed on a Cary 300 spectrophotometer (Agilent Technologies) (Mergny et al. 2005; Timmer et al. 2014).

IDS

Absorbance spectra from 220 to 335 nm for pG4 oligos were first measured in the absence of stabilizing cation. KCl was then added to each sample to a final concentration of 10 mM and absorbance spectra measured again after 30-min incubation at room temperature. IDS are presented as the arithmetic difference between the unfolded (no KCl) and folded (with 10 mM KCl) spectra. All spectra were measured at 37°C.

TDS

The folded G4 oligonucleotides in 10 mM KCl, 10 mM lithium cacodylate buffer (pH 7.0) were then used for TDS, determined using the arithmetic difference between the absorbance spectra from 220 to 335 nm at 95°C (unfolded) and 37°C (folded). Samples were held for 5 min at 95°C to ensure structures were fully unfolded before taking measurements. piM oligos were tested in 10 mM sodium cacodylate buffer (pH 5.5, 6.0, and 7.0) at 20°C (folded) and 90°C (unfolded).

UV melting

pG4 samples used in TDS experiment described above were slowly cooled from 95°C to 25°C at 0.5°C/min and absorbance at 260 nm taken every 1°C.

Circular dichroism

Strand concentrations of the sequences were adjusted to reach an absorbance of 0.8 at 260 nm according to their predicted extinction coefficient ε. pG4 samples were diluted in K10 buffer and piM samples in 10 mM sodium cacodylate (pH 5.0 and 6.0). Samples were heated at 65°C and allowed to cool to room temperature. The CD spectra were recorded at room temperature with three acquisitions per sample, 100 nm/min scanning speed, and 1.0-nm bandwidth.

Native PAGE

pG4 oligos (2 µM) in 50 mM KCl and 50 mM Tris (pH 7.0) were annealed by heating for 5 min at 90°C and then slowly cooling to room temperature for 2 h. A 20% (w/v) polyacrylamide gel was first prerun for 30 min at 90 V in 1 × TBE buffer with 10 mM KCl, after which samples were run at 90–100 V for ~3 h. After migration, the gel was incubated in a 0.5-µM ThT solution for 15 min with gentle agitation, protected from light, and visualized at λ_{ex} = 490 nm, λ_{em} = 520 nm. The gel was immediately counterstained with 1 × SYBR-safe in 1× TBE buffer for 40 min and visualized at λ_{ex} = 510 nm, λ_{em} = 530 nm.

Competing interest statement

The authors declare no competing interests.

Acknowledgments

C.W.E. acknowledges funding by the State Government of Western Australia. This study was funded by Agence Nationale de la Recherche (ANR) "G4Access" (ANR-20-CE12-0023), "iCare" (ANR-21-CE44-0005-01), and Institut National Du Cancer (INCa-PLBIO) "G4Access" grants to J.-L.M. N.M.S. acknowledges funding support from the National Health and Medical Research Council (NHMRC). Y.L. benefited from a Chinese Scholarship Council fellowship (201906340018).

Author contributions: The manuscript was written through contributions of all authors. All authors have given approval to the final version of the manuscript.

References

Abou Assi H, Garavís M, González C, Damha MJ. 2018. i-Motif DNA: structural features and significance to cell biology. *Nucleic Acids Res* **46:** 8038–8056. doi:10.1093/nar/gky735

Beknazarov N, Jin S, Poptsova M. 2020. Deep learning approach for predicting functional Z-DNA regions using omics data. *Sci Rep* **10:** 19134. doi:10.1038/s41598-020-76203-1

Belmonte-Reche E, Morales JC. 2020. G4-iM Grinder: when size and frequency matter. G-Quadruplex, i-Motif and higher order structure search and analysis tool. *NAR Genom Bioinform* **2:** lqz005. doi:10.1093/nargab/lqz005

Belmonte-Reche E, Serrano-Chacón I, Gonzalez C, Gallo J, Bañobre-López M. 2021. Potential G-quadruplexes and i-motifs in the SARS-CoV-2. *PLoS One* **16:** e0250654. doi:10.1371/journal.pone.0250654

Bhatt U, Kretzmann AL, Guédin A, Ou A, Kobelke S, Bond CS, Evans CW, Hurley LH, Mergny J-L, Iyer KS, et al. 2021. The role of G-quadruplex DNA in paraspeckle formation in cancer. *Biochimie* **190:** 124–131. doi:10.1016/j.biochi.2021.07.008

Biffi G, Tannahill D, Balasubramanian S. 2012. An intramolecular G-quadruplex structure is required for binding of telomeric repeat-containing RNA to the telomeric protein TRF2. *J Am Chem Soc* **134:** 11974–11976. doi:10.1021/ja305734x

Cabili MN, Trapnell C, Goff L, Koziol M, Tazon-Vega B, Regev A, Rinn JL. 2011. Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. *Genes Dev* **25:** 1915–1927. doi:10.1101/gad.17446611

Castillo-Guzman D, Chédin F. 2021. Defining R-loop classes and their contributions to genome instability. *DNA Repair (Amst)* **106:** 103182. doi:10.1016/j.dnarep.2021.103182

Chen L, Chen J-Y, Zhang X, Gu Y, Xiao R, Shao C, Tang P, Qian H, Luo D, Li H, et al. 2017. R-ChIP using inactive RNase H reveals dynamic coupling of R-loops with transcriptional pausing at gene promoters. *Mol Cell* **68:** 745–757.e5. doi:10.1016/j.molcel.2017.10.008

Chen J-Y, Zhang X, Fu X-D, Chen L. 2019. R-ChIP for genome-wide mapping of R-loops by using catalytically inactive RNASEH1. *Nat Protoc* **14:** 1661–1685. doi:10.1038/s41596-019-0154-6

Chillón I, Marcia M. 2020. The molecular structure of long non-coding RNAs: emerging patterns and functional implications. *Crit Rev Biochem Mol Biol* **55:** 662–690. doi:10.1080/10409238.2020.1828259

Collin D, Gehring K. 1998. Stability of chimeric DNA/RNA cytosine tetrads: implications for *i*-motif formation by RNA. *J Am Chem Soc* **120:** 4069–4072. doi:10.1021/ja973346r

Crossley MP, Bocek M, Cimprich KA. 2019. R-loops as cellular regulators and genomic threats. *Mol Cell* **73:** 398–411. doi:10.1016/j.molcel.2019.01.024

D'Ascenzo L, Leonarski F, Vicens Q, Auffinger P. 2016. 'Z-DNA like' fragments in RNA: a recurring structural motif with implications for folding, RNA/protein recognition and immune response. *Nucleic Acids Res* **44:** 5944–5956. doi:10.1093/nar/gkw388

De Cian A, Delemos E, Mergny J-L, Teulade-Fichou M-P, Monchaud D. 2007. Highly efficient G-quadruplex recognition by bisquinolinium compounds. *J Am Chem Soc* **129:** 1856–1857. doi:10.1021/ja067352b

de Reuver R, Dierick E, Wiernicki B, Staes K, Seys L, De Meester E, Muyldermans T, Botzki A, Lambrecht BN, Van Nieuwerburgh F, et al. 2021. ADAR1 interaction with Z-RNA promotes editing of endogenous double-stranded RNA and prevents MDA5-dependent immune activation. *Cell Rep* **36:** 109500. doi:10.1016/j.celrep.2021.109500

Derrien T, Johnson R, Bussotti G, Tanzer A, Djebali S, Tilgner H, Guernec G, Martin D, Merkel A, Knowles DG, et al. 2012. The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. *Genome Res* **22:** 1775–1789. doi:10.1101/gr.132159.111

Dumelie JG, Jaffrey SR. 2017. Defining the location of promoter-associated R-loops at near-nucleotide resolution using bisDRIP-seq. *eLife* **6:** e28306. doi:10.7554/eLife.28306

Fort V, Khelifi G, Hussein SMI. 2021. Long non-coding RNAs and transposable elements: a functional relationship. *Biochim Biophys Acta* **1868:** 118837. doi:10.1016/j.bbamcr.2020.118837

Frankish A, Diekhans M, Ferreira A-M, Johnson R, Jungreis I, Loveland J, Mudge JM, Sisu C, Wright J, Armstrong J, et al. 2019. GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Res* **47:** D766–D773. doi:10.1093/nar/gky955

Frasson I, Pirota V, Richter SN, Doria F. 2022. Multimeric G-quadruplexes: a review on their biological roles and targeting. *Int J Biol Macromol* **204:** 89–102. doi:10.1016/j.ijbiomac.2022.01.197

Fukuhara M, Ma Y, Nagasawa K, Toyoshima F. 2017. A G-quadruplex structure at the 5′ end of the *H19* coding region regulates *H19* transcription. *Sci Rep* **7:** 45815. doi:10.1038/srep45815

Garant JM, Perreault JP, Scott MS. 2017. Motif independent identification of potential RNA G-quadruplexes by G4RNA screener. *Bioinformatics* **33:** 3532–3537. doi:10.1093/bioinformatics/btx498

Gaspar P, Moura G, Santos MAS, Oliveira JL. 2013. mRNA secondary structure optimization using a correlated stem–loop prediction. *Nucleic Acids Res* **41:** e73. doi:10.1093/nar/gks1473

Ghafouri-Fard S, Abak A, Baniahmad A, Hussen BM, Taheri M, Jamali E, Dinger ME. 2022. Interaction between non-coding RNAs, mRNAs and G-quadruplexes. *Cancer Cell Int* **22:** 171. doi:10.1186/s12935-022-02601-2

Ghosh A, Pandey SP, Ansari AH, Sundar JS, Singh P, Khan Y, Ekka MK, Chakraborty D, Maiti S. 2022. Alternative splicing modulation mediated by G-quadruplex structures in MALAT1 lncRNA. *Nucleic Acids Res* **50:** 378–396. doi:10.1093/nar/gkab1066

Graf J, Kretz M. 2020. From structure to function: route to understanding lncRNA mechanism. *Bioessays* **42:** 2000027. doi:10.1002/bies.202000027

Guédin A, Lin LY, Armane S, Lacroix L, Mergny J-L, Thore S, Yatsunyk LA. 2018. Quadruplexes in '*Dicty*': crystal structure of a four-quartet G-quadruplex formed by G-rich motif found in the *Dictyostelium discoideum* genome. *Nucleic Acids Res* **46:** 5297–5307. doi:10.1093/nar/gky290

Guo JU, Bartel DP. 2016. RNA G-quadruplexes are globally unfolded in eukaryotic cells and depleted in bacteria. *Science* **353:** aaf5371. doi:10.1126/science.aaf5371

Herbert A. 2019. Z-DNA and Z-RNA in human disease. *Commun Biol* **2:** 7. doi:10.1038/s42003-018-0237-x

Herrera-Solorio AM, Peralta-Arrieta I, Armas López L, Hernández-Cigala N, Mendoza Milla C, Ortiz Quintero B, Catalán Cárdenas R, Pineda Villegas P, Rodríguez Villanueva E, Trejo Iriarte CG, et al. 2021. LncRNA SOX2-OT regulates AKT/ERK and SOX2/GLI-1 expression, hinders therapy, and worsens clinical prognosis in malignant lung diseases. *Mol Oncol* **15:** 1110–1129. doi:10.1002/1878-0261.12875

Ho PS, Ellison MJ, Quigley GJ, Rich A. 1986. A computer aided thermodynamic approach for predicting the formation of Z-DNA in naturally occurring sequences. *EMBO J* **5:** 2737–2744. doi:10.1002/j.1460-2075.1986.tb04558.x

Hon C-C, Ramilowski JA, Harshbarger J, Bertin N, Rackham OJL, Gough J, Denisenko E, Poulsen TM, Severin J, Lizio M, et al. 2017. An atlas of human long non-coding RNAs with accurate 5′ ends. *Nature* **543:** 199–204. doi:10.1038/nature21374

Hou Z, Zhao W, Zhou J, Shen L, Zhan P, Xu C, Chang C, Bi H, Zou J, Yao X, et al. 2014. A long noncoding RNA Sox2ot regulates lung cancer cell proliferation and is a prognostic indicator of poor survival. *Int J Biochem Cell Biol* **53:** 380–388. doi:10.1016/j.biocel.2014.06.004

Huang T, Wang Y-J, Huang M-T, Guo Y, Yang L-C, Liu X-J, Tan W-Y, Long J-H. 2021. LINC00470 accelerates the proliferation and metastasis of melanoma through promoting APEX1 expression. *Cell Death Dis* **12:** 410. doi:10.1038/s41419-021-03612-z

Irving KL, King JJ, Waller ZAE, Evans CW, Smith NM. 2022. Stability and context of intercalated motifs (i-motifs) for biological applications. *Biochimie* **198:** 33–47. doi:10.1016/j.biochi.2022.03.001

Jain A, Vale RD. 2017. RNA phase transitions in repeat expansion disorders. *Nature* **546:** 243–247. doi:10.1038/nature22386

Jayaraj GG, Pandey S, Scaria V, Maiti S. 2012. Potential G-quadruplexes in the human long non-coding transcriptome. *RNA Biol* **9:** 81–89. doi:10.4161/rna.9.1.18047

Jenjaroenpun P, Wongsurawat T, Yenamandra SP, Kuznetsov VA. 2015. QmRLFS-finder: a model, web server and stand-alone tool for prediction and analysis of R-loop forming sequences. *Nucleic Acids Res* **43:** W527–W534. doi:10.1093/nar/gkv344

Jiao H, Wachsmuth L, Wolf S, Lohmann J, Nagata M, Kaya GG, Oikonomou N, Kondylis V, Rogg M, Diebold M, et al. 2022. ADAR1 averts fatal type I interferon induction by ZBP1. *Nature* **607:** 776–783. doi:10.1038/s41586-022-04878-9

Johnson R, Guigó R. 2014. The RIDL hypothesis: transposable elements as functional domains of long noncoding RNAs. *RNA* **20:** 959–976. doi:10.1261/rna.044560.114

Johnsson P, Lipovich L, Grandér D, Morris KV. 2014. Evolutionary conservation of long non-coding RNAs; sequence, structure, function. *Biochim Biophys Acta* **1840:** 1063–1071. doi:10.1016/j.bbagen.2013.10.035

Kapusta A, Kronenberg Z, Lynch VJ, Zhuo X, Ramsay L, Bourque G, Yandell M, Feschotte C. 2013. Transposable elements are major contributors to the origin, diversification, and regulation of vertebrate long noncoding RNAs. *PLoS Genet* **9:** e1003470. doi:10.1371/journal.pgen.1003470

Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D. 2002. The human genome browser at UCSC. *Genome Res* **12:** 996–1006. doi:10.1101/gr.229102

Knauss JL, Miao N, Kim S-N, Nie Y, Shi Y, Wu T, Pinto HB, Donohoe ME, Sun T. 2018. Long noncoding RNA *Sox2ot* and transcription factor YY1 co-regulate the differentiation of cortical neural progenitors by repressing *Sox2*. *Cell Death Dis* **9:** 799. doi:10.1038/s41419-018-0840-2

Koeris M, Funke L, Shrestha J, Rich A, Maas S. 2005. Modulation of ADAR1 editing activity by Z-RNA in vitro. *Nucleic Acids Res* **33:** 5362–5370. doi:10.1093/nar/gki849

Kwok CK, Marsico G, Sahakyan AB, Chambers VS, Balasubramanian S. 2016. rG4-seq reveals widespread formation of G-quadruplex structures in the human transcriptome. *Nat Methods* **13:** 841–844. doi:10.1038/nmeth.3965

Kypr J, Kejnovská I, Renčiuk D, Vorlíčková M. 2009. Circular dichroism and conformational polymorphism of DNA. *Nucleic Acids Res* **37:** 1713–1725. doi:10.1093/nar/gkp026

Lacroix L, Mergny J-L, Leroy J-L, Hélène C. 1996. Inability of RNA To form the i-motif: implications for triplex formation. *Biochemistry* **35:** 8715–8722. doi:10.1021/bi960107s

Lee A-R, Hwang J, Hur JH, Ryu K-S, Kim KK, Choi B-S, Kim N-K, Lee J-H. 2019. NMR dynamics study reveals the Zα domain of human ADAR1 associates with and dissociates from Z-RNA more slowly than Z-DNA. *ACS Chem Biol* **14:** 245–255. doi:10.1021/acschembio.8b00914

Lexa M, Steflova P, Martinek T, Vorlickova M, Vyskot B, Kejnovsky E. 2014. Guanine quadruplexes are formed by specific regions of human transposable elements. *BMC Genomics* **15:** 1032. doi:10.1186/1471-2164-15-1032

Lin R, Zhong X, Zhou Y, Geng H, Hu Q, Huang Z, Hu J, Fu X-D, Chen L, Chen J-Y. 2022. R-loopBase: a knowledgebase for genome-wide R-loop formation and regulation. *Nucleic Acids Res* **50:** D303–D315. doi:10.1093/nar/gkab1103

Liu C, Fu H, Liu X, Lei Q, Zhang Y, She X, Liu Q, Sun Y, Li G, Wu M. 2018a. LINC00470 coordinates the epigenetic regulation of ELFN2 to distract GBM cell autophagy. *Mol Ther* **26:** 2267–2281. doi:10.1016/j.ymthe.2018.06.019

Liu C, Zhang Y, She X, Fan L, Li P, Feng J, Fu H, Liu Q, Liu Q, Zhao C, et al. 2018b. A cytoplasmic long noncoding RNA LINC00470 as a new AKT activator to mediate glioblastoma cell autophagy. *J Hematol Oncol* **11:** 77. doi:10.1186/s13045-018-0619-z

Liu Y, Wu Z, Zhou J, Ramadurai DKA, Mortenson KL, Aguilera-Jimenez E, Yan Y, Yang X, Taylor AM, Varley KE, et al. 2021. A predominant enhancer co-amplified with the SOX2 oncogene is necessary and sufficient for its expression in squamous cancer. *Nat Commun* **12:** 7139. doi:10.1038/s41467-021-27055-4

Luo Y, Granzhan A, Verga D, Mergny J-L. 2021. FRET-MC: a fluorescence melting competition assay for studying G4 structures in vitro. *Biopolymers* **112:** e23415. doi:10.1002/bip.23415

Lv D, Xu K, Jin X, Li J, Shi Y, Zhang M, Jin X, Li Y, Xu J, Li X. 2020. LncSpA: LncRNA spatial atlas of expression across normal and cancer tissues. *Cancer Res* **80:** 2067–2071. doi:10.1158/0008-5472.CAN-19-2687

Lyu K, Chow EY-C, Mou X, Chan T-F, Kwok CK. 2021. RNA G-quadruplexes (rG4s): genomics and biological functions. *Nucleic Acids Res* **49:** 5426–5450. doi:10.1093/nar/gkab187

Lyu J, Shao R, Kwong Yung PY, Elsässer SJ. 2022. Genome-wide mapping of G-quadruplex structures with CUT&Tag. *Nucleic Acids Res* **50:** e13. doi:10.1093/nar/gkab1073

Makova KD, Weissensteiner MH. 2023. Noncanonical DNA structures are drivers of genome evolution. *Trends Genet* **39:** 109–124. doi:10.1016/j.tig.2022.11.005

Malig M, Hartono SR, Giafaglione JM, Sanz LA, Chedin F. 2020. Ultra-deep coverage single-molecule R-loop footprinting reveals principles of R-loop formation. *J Mol Biol* **432:** 2271–2288. doi:10.1016/j.jmb.2020.02.014

Matsumura K, Kawasaki Y, Miyamoto M, Kamoshida Y, Nakamura J, Negishi L, Suda S, Akiyama T. 2017. The novel G-quadruplex-containing long non-coding RNA GSEC antagonizes DHX36 and modulates colon cancer cell migration. *Oncogene* **36:** 1191–1199. doi:10.1038/onc.2016.282

Mattick JS, Amaral PP, Carninci P, Carpenter S, Chang HY, Chen L-L, Chen R, Dean C, Dinger ME, Fitzgerald KA, et al. 2023. Long non-coding RNAs: definitions, functions, challenges and recommendations. *Nat Rev Mol Cell Biol* **24:** 430–447. doi:10.1038/s41580-022-00566-8

Mauger DM, Cabral BJ, Presnyak V, Su SV, Reid DW, Goodman B, Link K, Khatwani N, Reynders J, Moore MJ, et al. 2019. mRNA structure regulates protein expression through changes in functional half-life. *Proc Natl Acad Sci* **116:** 24075–24083. doi:10.1073/pnas.1908052116

McBrayer D, Schoonover M, Long KJ, Escobedo R, Kerwin SM. 2019. *N*-Methylmesoporphyrin IX exhibits G-quadruplex-specific photocleavage activity. *Chembiochem* **20:** 1924–1927. doi:10.1002/cbic.201900002

Mergny J-L, Li J, Lacroix L, Amrane S, Chaires JB. 2005. Thermal difference spectra: a specific signature for nucleic acid structures. *Nucleic Acids Res* **33:** e138. doi:10.1093/nar/gni134

Messemaker TC, van Leeuwen SM, van den Berg PR, 't Jong AEJ, Palstra R-J, Hoeben RC, Semrau S, Mikkers HMM. 2018. Allele-specific repression of *Sox2* through the long non-coding RNA *Sox2ot*. *Sci Rep* **8:** 386. doi:10.1038/s41598-017-18649-4

Mirzaei S, Paskeh MDA, Entezari M, Mirmazloomi SR, Hassanpoor A, Aboutalebi M, Rezaei S, Hejazi ES, Kakavand A, Heidari H, et al. 2022. SOX2 function in cancers: association with growth, invasion, stemness and therapy response. *Biomed Pharmacother* **156:** 113860. doi:10.1016/j.biopha.2022.113860

Miyahara T, Nakatsuji H, Sugiyama H. 2016. Similarities and differences between RNA and DNA double-helical structures in circular dichroism spectroscopy: a SAC–CI study. *J Phys Chem A* **120:** 9008–9018. doi:10.1021/acs.jpca.6b08023

Nakahama T, Kawahara Y. 2021. Deciphering the biological significance of ADAR1–Z-RNA interactions. *Int J Mol Sci* **22:** 11435. doi:10.3390/ijms222111435

Nakahama T, Kato Y, Shibuya T, Inoue M, Kim JI, Vongpipatana T, Todo H, Xing Y, Kawahara Y. 2021. Mutations in the adenosine deaminase ADAR1 that prevent endogenous Z-RNA binding induce Aicardi-Goutières-syndrome-like encephalopathy. *Immunity* **54:** 1976–1988.e7. doi:10.1016/j.immuni.2021.08.022

Ou M, Li X, Zhao S, Cui S, Tu J. 2020. Long non-coding RNA CDKN2B-AS1 contributes to atherosclerotic plaque formation by forming RNA-DNA triplex in the CDKN2B promoter. *EBioMedicine* **55:** 102694. doi:10.1016/j.ebiom.2020.102694

Paramasivan S, Rujan I, Bolton PH. 2007. Circular dichroism of quadruplex DNAs: applications to structure, cation effects and ligand binding. *Methods* **43:** 324–331. doi:10.1016/j.ymeth.2007.02.009

Pegueroles C, Gabaldón T. 2016. Secondary structure impacts patterns of selection in human lncRNAs. *BMC Biol* **14:** 60. doi:10.1186/s12915-016-0283-0

Petermann E, Lan L, Zou L. 2022. Sources, resolution and physiological relevance of R-loops and RNA–DNA hybrids. *Nat Rev Mol Cell Biol* **23:** 521–540. doi:10.1038/s41580-022-00474-x

Pintacuda G, Young AN, Cerase A. 2017. Function by structure: spotlights on xist long non-coding RNA. *Front Mol Biosci* **4:** 90. doi:10.3389/fmolb.2017.00090

Placido D, Brown BA, Lowenhaupt K, Rich A, Athanasiadis A. 2007. A left handed RNA double helix bound by the Zα domain of the RNA editing enzyme ADAR1. *Structure* **15:** 395–404. doi:10.1016/j.str.2007.03.001

R Core Team. 2023. *R: a language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna. https://www.R-project.org/.

Saccà B, Lacroix L, Mergny J-L. 2005. The effect of chemical modifications on the thermal stability of different G-quadruplex-forming oligonucleotides. *Nucleic Acids Res* **33:** 1182–1192. doi:10.1093/nar/gki257

Sahakyan AB, Chambers VS, Marsico G, Santner T, Di Antonio M, Balasubramanian S. 2017a. Machine learning model for sequence-driven DNA G-quadruplex formation. *Sci Rep* **7:** 14535. doi:10.1038/s41598-017-14017-4

Sahakyan AB, Murat P, Mayer C, Balasubramanian S. 2017b. G-quadruplex structures within the 3′ UTR of LINE-1 elements stimulate retrotransposition. *Nat Struct Mol Biol* **24:** 243–247. doi:10.1038/nsmb.3367

Sanz LA, Chédin F. 2019. High-resolution, strand-specific R-loop mapping via S9.6-based DNA–RNA immunoprecipitation and high-throughput sequencing. *Nat Protoc* **14:** 1734–1755. doi:10.1038/s41596-019-0159-1

Schaefer T, Lengerke C. 2020. SOX2 protein biochemistry in stemness, reprogramming, and cancer: the PI3K/AKT/SOX2 axis and beyond. *Oncogene* **39:** 278–292. doi:10.1038/s41388-019-0997-x

Schlackow M, Nojima T, Gomes T, Dhir A, Carmo-Fonseca M, Proudfoot NJ. 2017. Distinctive patterns of transcription and RNA processing for human lincRNAs. *Mol Cell* **65:** 25–38. doi:10.1016/j.molcel.2016.11.029

Seemann SE, Mirza AH, Hansen C, Bang-Berthelsen CH, Garde C, Christensen-Dalsgaard M, Torarinsson E, Yao Z, Workman CT, Pociot F, et al. 2017. The identification and functional annotation of RNA

structures conserved in vertebrates. *Genome Res* **27:** 1371–1383. doi:10.1101/gr.208652.116
Shafiee M, Aleyasin SA, Vasei M, Semnani SS, Mowla SJ. 2016. Down-regulatory effects of miR-211 on long non-coding RNA SOX2OT and SOX2 genes in esophageal squamous cell carcinoma. *Cell J* **17:** 593–600. doi:10.22074/cellj.2016.3811
Shahryari A, Jazi MS, Samaei NM, Mowla SJ. 2015. Long non-coding RNA SOX2OT: expression signature, splicing patterns, and emerging roles in pluripotency and tumorigenesis. *Front Genet* **6:** 196. doi:10.3389/fgene.2015.00196
Simko EAJ, Liu H, Zhang T, Velasquez A, Teli S, Haeusler AR, Wang J. 2020. G-quadruplexes offer a conserved structural motif for NONO recruitment to NEAT1 architectural lncRNA. *Nucleic Acids Res* **48:** 7421–7438. doi:10.1093/nar/gkaa475
Simone R, Fratta P, Neidle S, Parkinson GN, Isaacs AM. 2015. G-quadruplexes: emerging roles in neurodegenerative diseases and the non-coding transcriptome. *FEBS Lett* **589:** 1653–1668. doi:10.1016/j.febslet.2015.05.003
Smit AFA, Hubley R, Green P. 2015. RepeatMasker Open-4.0. Institute for Systems Biology. https://www.repeatmasker.org/RepeatMasker/.
Snoussi K, Nonin-Lecomte S, Leroy J-L. 2001. The RNA i-motif. *J Mol Biol* **309:** 139–153. doi:10.1006/jmbi.2001.4618
Statello L, Guo C-J, Chen L-L, Huarte M. 2021. Gene regulation by long non-coding RNAs and its biological functions. *Nat Rev Mol Cell Biol* **22:** 96–118. doi:10.1038/s41580-020-00315-9
Su R, Cao S, Ma J, Liu Y, Liu X, Zheng J, Chen J, Liu L, Cai H, Li Z, et al. 2017. Knockdown of SOX2OT inhibits the malignant biological behaviors of glioblastoma stem cells via up-regulating the expression of miR-194-5p and miR-122. *Mol Cancer* **16:** 171. doi:10.1186/s12943-017-0737-1
Tang Q, Rigby RE, Young GR, Hvidt AK, Davis T, Tan TK, Bridgeman A, Townsend AR, Kassiotis G, Rehwinkel J. 2021. Adenosine-to-inosine editing of endogenous Z-form RNA by the deaminase ADAR1 prevents spontaneous MAVS-dependent type I interferon responses. *Immunity* **54:** 1961–1975.e5. doi:10.1016/j.immuni.2021.08.011
Tassinari M, Richter SN, Gandellini P. 2021. Biological relevance and therapeutic potential of G-quadruplex structures in the human noncoding transcriptome. *Nucleic Acids Res* **49:** 3617–3633. doi:10.1093/nar/gkab127
Tateishi-Karimata H, Sugimoto N. 2021. Roles of non-canonical structures of nucleic acids in cancer and neurodegenerative diseases. *Nucleic Acids Res* **49:** 7839–7855. doi:10.1093/nar/gkab580
Tavares RCA, Pyle AM, Somarowthu S. 2019. Phylogenetic analysis with improved parameters reveals conservation in lncRNA structures. *J Mol Biol* **431:** 1592–1603. doi:10.1016/j.jmb.2019.03.012
Teng Y, Kang H, Chu Y. 2019. Identification of an exosomal long noncoding RNA SOX2-OT in plasma as a promising biomarker for lung squamous cell carcinoma. *Genet Test Mol Biomarkers* **23:** 235–240. doi:10.1089/gtmb.2018.0103
Timmer CM, Michmerhuizen NL, Witte AB, Van Winkle M, Zhou D, Sinniah K. 2014. An isothermal titration and differential scanning calorimetry study of the G-quadruplex DNA–insulin interaction. *J Phys Chem B* **118:** 1784–1790. doi:10.1021/jp411293r
Uroda T, Anastasakou E, Rossi A, Teulon J-M, Pellequer J-L, Annibale P, Pessey O, Inga A, Chillón I, Marcia M. 2019. Conserved pseudoknots in lncRNA MEG3 are essential for stimulation of the p53 pathway. *Mol Cell* **75:** 982–995.e9. doi:10.1016/j.molcel.2019.07.025
Varshney D, Spiegel J, Zyner K, Tannahill D, Balasubramanian S. 2020. The regulation and functions of DNA and RNA G-quadruplexes. *Nat Rev Mol Cell Biol* **21:** 459–474. doi:10.1038/s41580-020-0236-x
Varshney D, Cuesta SM, Herdy B, Abdullah UB, Tannahill D, Balasubramanian S. 2021. RNA G-quadruplex structures control ribosomal protein production. *Sci Rep* **11:** 22735. doi:10.1038/s41598-021-01847-6
Wahba L, Costantino L, Tan FJ, Zimmer A, Koshland D. 2016. S1-DRIP-seq identifies high expression and polyA tracts as major contributors to R-loop formation. *Genes Dev* **30:** 1327–1338. doi:10.1101/gad.280834.116
Wang AH, Gessner RV, van der Marel GA, van Boom JH, Rich A. 1985. Crystal structure of Z-DNA without an alternating purine-pyrimidine sequence. *Proc Natl Acad Sci* **82:** 3611–3615. doi:10.1073/pnas.82.11.3611
Wang E, Thombre R, Shah Y, Latanich R, Wang J. 2021a. G-Quadruplexes as pathogenic drivers in neurodegenerative disorders. *Nucleic Acids Res* **49:** 4816–4830. doi:10.1093/nar/gkab164
Wang K, Wang H, Li C, Yin Z, Xiao R, Li Q, Xiang Y, Wang W, Huang J, Chen L, et al. 2021b. Genomic profiling of native R loops with a DNA-RNA hybrid recognition sensor. *Sci Adv* **7:** eabe3516. doi:10.1126/sciadv.abe3516
Wu T, Hu E, Xu S, Chen M, Guo P, Dai Z, Feng T, Zhou L, Tang W, Zhan L, et al. 2021. clusterProfiler 4.0: a universal enrichment tool for interpreting omics data. *Innovation (Camb)* **2:** 100141. doi:10.1016/j.xinn.2021.100141
Xu S, Li Q, Xiang J, Yang Q, Sun H, Guan A, Wang L, Liu Y, Yu L, Shi Y, et al. 2016. Thioflavin T as an efficient fluorescence sensor for selective recognition of RNA G-quadruplexes. *Sci Rep* **6:** 24793. doi:10.1038/srep24793
Yan J, Huang X, Zhang X, Chen Z, Ye C, Xiang W, Huang Z. 2020. LncRNA LINC00470 promotes the degradation of PTEN mRNA to facilitate malignant behavior in gastric cancer cells. *Biochem Biophys Res Commun* **521:** 887–893. doi:10.1016/j.bbrc.2019.11.016
Yang SY, Lejault P, Chevrier S, Boidot R, Robertson AG, Wong JMY, Monchaud D. 2018. Transcriptome-wide identification of transient RNA G-quadruplexes in human cells. *Nat Commun* **9:** 4730. doi:10.1038/s41467-018-07224-8
Yang Z, Li M, Sun Q. 2020. RHON1 Co-transcriptionally resolves R-loops for *Arabidopsis* chloroplast genome maintenance. *Cell Rep* **30:** 243–256.e5. doi:10.1016/j.celrep.2019.12.007
Yang SY, Monchaud D, Wong JMY. 2022. Global mapping of RNA G-quadruplexes (G4-RNAs) using G4RP-seq. *Nat Protoc* **17:** 870–889. doi:10.1038/s41596-021-00671-6
Yi T, Song Y, Zuo L, Wang S, Miao J. 2021. LINC00470 stimulates methylation of PTEN to facilitate the progression of endometrial cancer by recruiting DNMT3a through MYC. *Front Oncol* **11:** 646217. doi:10.3389/fonc.2021.646217
Zaccaria F, Fonseca Guerra C. 2018. RNA versus DNA G-quadruplex: the origin of increased stability. *Chemistry* **24:** 16315–16322. doi:10.1002/chem.201803530
Zampetaki A, Albrecht A, Steinhofel K. 2018. Long non-coding RNA structure and function: is there a link? *Front Physiol* **9:** 1201. doi:10.3389/fphys.2018.01201
Zeraati M, Langley DB, Schofield P, Moye AL, Rouet R, Hughes WE, Bryan TM, Dinger ME, Christ D. 2018. I-motif DNA structures are formed in the nuclei of human cells. *Nat Chem* **10:** 631–637. doi:10.1038/s41557-018-0046-3
Zhan Y, Chen Z, He S, Gong Y, He A, Li Y, Zhang L, Zhang X, Fang D, Li X, et al. 2020. Long non-coding RNA SOX2OT promotes the stemness phenotype of bladder cancer cells by modulating SOX2. *Mol Cancer* **19:** 25. doi:10.1186/s12943-020-1143-7
Zhang J-J, Zhu Y, Zhang X-F, Liu D-F, Wang Y, Yang C, Shi G-D, Peng Y-P, Zhang K, Tian L, et al. 2017. Yin yang-1 suppresses pancreatic ductal adenocarcinoma cell proliferation and tumor growth by regulating SOX2OT-SOX2 axis. *Cancer Lett* **408:** 144–154. doi:10.1016/j.canlet.2017.08.032
Zhang W, Yang S, Chen D, Yuwen D, Zhang J, Wei X, Han X, Guan X. 2022. SOX2-OT induced by PAI-1 promotes triple-negative breast cancer cells metastasis by sponging miR-942-5p and activating PI3K/Akt signaling. *Cell Mol Life Sci* **79:** 59. doi:10.1007/s00018-021-04120-1
Zillinger T, Bartok E. 2021. ADAR1 edits the SenZ and SenZ-ability of RNA. *Immunity* **54:** 1909–1911. doi:10.1016/j.immuni.2021.08.021
Zyner KG, Simeone A, Flynn SM, Doyle C, Marsico G, Adhikari S, Portella G, Tannahill D, Balasubramanian S. 2022. G-quadruplex DNA structures in human stem cells and differentiation. *Nat Commun* **13:** 142. doi:10.1038/s41467-021-27719-1

Received June 30, 2023; accepted in revised form January 31, 2024.