Biochemistry

pubs.acs.org/biochemistry Article

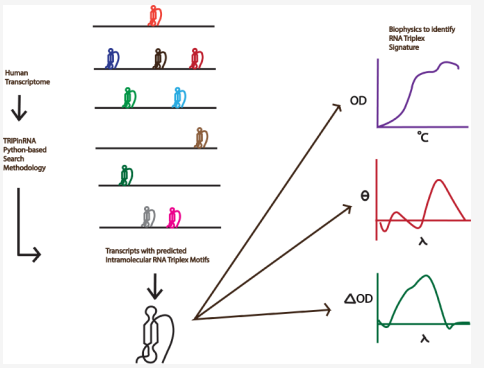
Development of an In Silico Platform (TRIPinRNA) for the Identification of Novel RNA Intramolecular Triple Helices and Their Validation Using Biophysical Techniques

Published as part of Biochemistry special issue "Computational Biochemistry".

Isha Rakheja,[‖] Vishal Bharti,[‖] S Sahana, Prosad Kumar Das, Gyan Ranjan, Ajit Kumar, Niyati Jain, and Souvik Maiti*

Cite This: *Biochemistry* 2025, 64, 250−265

ABSTRACT: There are surprisingly few RNA intramolecular triple helices known in the human transcriptome. The structure has been most well-studied as a stability-element at the 3′ end of lncRNAs such as *MALAT1* and *NEAT1*, but the intrigue remains whether it is indeed as rare as it is understood to be or just waiting for a closer look from a new vantage point. TRIPinRNA, our Python-based in silico platform, allows for a comprehensive sequence-pattern search for potential triplex formation in the human transcriptome—noncoding as well as coding. Using this tool, we report the putative occurrence of homopyrimidine type (canonical) triple helices as well as heteropurine−pyrimidine strand type (noncanonical) triple helices in the human transcriptome and validate the formation of both types of triplexes using biophysical approaches. We find that the occurrence of triplex structures has a strong correlation with local GC content, which might be influencing their formation. By employing a search that encompasses both canonical and noncanonical triplex structures across the human transcriptome, this study enriches the understanding of RNA biology. Lastly, TRIPinRNA can be utilized in finding triplex structures for any organism with an annotated transcriptome.



INTRODUCTION

Noncoding RNAs (lncRNAs) being single stranded have a greater propensity than DNA to fold into secondary and tertiary structures. They are known to be structurally conserved and employ distinct motifs to assist in their function.[1−3] Several canonical and noncanonical structures in RNA molecules in general have been shown to serve important functions, for example, in the tRNA cloverleaf structures, the rRNA scaffolds, G-quadruplex structures at telomere ends, pseudoknots, and stability-element-type triple helix structures at the 3′ ends of lncRNAs like *MALAT1* and *NEAT1*.[4−12]

The formation of the unique stability-element type RNA intramolecular triple helix structure is one of the different cellular mechanisms in place for preventing RNA degradation.[13,14] This triple helix structure is relatively well-studied but has been reported for a very limited number of genes.[6,15,16] Among these, *MALAT1* and *NEAT1* (*MENß*) and the viral KSHV *PAN* RNA lncRNAs contain pyrimidine-type triple helices at their 3′ ends.[17] In addition, the human telomerase RNA (hTR) is known to form a triple helix with a short stretch of three triplets of U.A-U within its pseudoknot region,[5,18,19] and in the SAM-II riboswitch of bacteria, the formation of a 5-base triple helix structure is the basis on which the switching function proceeds.[20]

In recent times, due to the increase in efforts to therapeutically target human RNAs, the use of small molecules as robust perturbation agents has gained limelight.[21−24] In the context of a triple helix structure, the binding of such molecules may lead to disruption of the said stability element structure, in effect controlling the level of the associated RNA by drug dosing.[25,26] This has sparked interest in the search for more *MALAT1*-like intramolecular triple helix motifs. Along these lines, in fungi and plants, the presence of dENEs (double domain ENE) has been reported by Steitz et al. by using Infernal, an RNA homology search tool.[27−30] Another seminal report investigated for the presence of intramolecular RNA triple helix motifs across various organisms using a structure-based search.[31] Here, the authors found the presence of 44 such loci in the lizard *Anolis carolinensis*, 34 in medaka, and 8 in zebrafish. However, in the

Received: June 14, 2024
Revised: November 25, 2024
Accepted: December 5, 2024
Published: December 13, 2024

https://doi.org/10.1021/acs.biochem.4c00334
Biochemistry 2025, 64, 250−265

human transcriptome, their search again yielded just the two previously known triple helices of *MALAT1* and *NEAT1*.

There are a few other programs capable of searching for intramolecular triple helix sequences. However, these contain limitations in the length of gap sequence permitted (a maximum of 10 nucleotides) between the Hoogsteen−Crick and the Crick−Watson strand.[32−36] This would be ineffective in searching for the occurrence of *MALAT1*-like triple helices. The significance of loop length toward the stability of RNA structures has been previously discussed and from common understanding, we know that the entropy lost on having a larger loop between triple helix strands could be compensated by an enthalpy gain in cases where a stem is predicted to form in that sequence.[13,37,38] To the best of our knowledge, only one program—NeSSie[39]—allows longer gaps, but due to its lenient search criteria is not capable of yielding a workable number of results for *MALAT1*-like triplexes.

The question then still remains—are triple helix structures so rare that in the close to 200,000 noncoding RNA sequences[40] reported in the human transcriptome (NONCODE), essentially only two intramolecular triple helices are known? It is conceivable that there is still scope for exploration regarding triple helix forming sequences in the human transcriptome, and to this end, we report the development of a computational platform called TRIPinRNA that fills this lacuna by employing a sequence pattern-based search. Using the two well-known triple helix sequences in the human transcriptome, MALAT1 and NEAT1- as templates, we comprehensively study the GENCODE human lncRNA sequence data set for parallel-type triple helices. Additionally, we postulated that intramolecular triple helix structures may be present at locations other than 3′ ends in the RNA transcripts in order to be performing other as-yet unreported functions (similar to the functions of other RNA structures), and hence, we have searched the complete transcript sequences. Further, the results obtained from the human lncRNA data set have been compared to our search results on the human protein coding transcriptome. Finally, the obtained sequences have been validated by using biophysical techniques to form triplexes.

Taken together, we report the presence of plausible triplex-forming sequences in the lncRNA genes for X-chromosome inactivation. In the coding transcriptome, we find that the maximum number of hits fall in the 3′ UTR regions and that the 5′ UTR and 3′ UTR regions of transcripts with hits display a preference for certain GC content. This correlation is strikingly absent from the CDS regions for these transcripts when compared to the GC usage expected on average. This exploration broadens our comprehension of RNA structural biology and underscores the future therapeutic potential of these noncanonical triple helices.

MATERIALS AND METHODS

Writing a Python Script for TRIPinRNA: Experimental Procedure and Analysis Workflow—Bioinformatics Algorithm. A Python-based script was designed to analyze the entire transcriptome, encompassing both coding and long noncoding sequences. The underlying algorithm employed a string match search. Scanning was performed for core triplex structures that comprised three strands: the Hoogsteen strand (H strand), the Watson strand (W strand), and the Crick strand (C strand). A positive hit was identified when the script detected these strands with specific constraints: a complementary base pairing between the W and C strands and H and W strands, and a unique feature where the H strand is one nucleotide shorter than the other two strands (for search "with-gap"). This "with-gap" configuration search is contrasted in another script variant that does not account for such a gap (search "without-gap").

A sliding window search mechanism is employed that varies in length across each transcript, seeking triple helix structures within these defined parameters. The permissible length of the Hoogsteen (H) strand may range between 7 and 15 nucleotides (nt). Hoogsteen base-pairing is expected to occur between the H strand and the W strand. (Only canonical base-pairing is taken into account, and noncanonical base-pairing is not considered anywhere in this script.) The gap in the H strand may be positioned variably, with the exception of 3 nucleotides (nt) at each end of the H strand based on criteria defined by us. Constraints also include a minimum of 3 bp length for the upper stem (if found) and a minimum 3 nt loop of this stem (mandatory minimum criteria), with a combined maximum of 100 nt (and a minimum length of 3 nt) for this upper sequence between H and C strands, regardless of whether it contains a stem. Similar constraints apply in the intervening sequence between C and W strands, and if no stem is detected within the 3−100 nt range, the sequence is classified as a lower loop.

The algorithm first searches for the core H−W−C triplex structure, followed by assessing the constraints for the upper and lower stem-loops. A positive hit is reported only after these criteria are met. Our script exhaustively checks all possible combinations within the allowed nucleotide range and gap positions in the H strand, reporting all identified potential triple helix structures in the transcriptome. The search commences with an initial padding of 3 nt at the start of any transcript, followed by a continuous sliding window approach. An alternate version of the script, which omits the gap in the H strand (termed as "without gap" or a case with "no gap"), follows a similar algorithm.

The script takes a zipped FASTA file as input, and the output is a CSV format file with all of the nucleotide sequences and the start and the end indices for all of the features detected in the reported triplex structure.

This versatile script is adaptable for analyzing the transcriptome of various organisms or species.

High-Performance Computing in Parallel Processing. To efficiently handle the computational demands of this analysis, a high-performance computing (HPC) environment was employed. The scripts were executed using a Bash script on an HPC cluster, leveraging the SLURM workload manager for job scheduling and resource management.

Furthermore, to promote the reproducibility of our study and to allow for community verification and extension, the source code was made available for both versions of our script in the GitHub repository. The code may be found at GitHub URL: https://github.com/visvikbharti/TripInRNA.

Bioinformatics: Downstream Analysis Workflow. Both versions of the script were run for two types of transcriptomes: the protein coding-based human transcriptome (using the FASTA file gencode.v44.pc_transcripts.fa.gz) and the lncRNA-based human transcriptome (using the FASTA file gencode.v44.lncRNA_transcripts.fa.gz). Both were downloaded from the GENCODE database (release 44 for human).

Statistical Analysis of the GC/AU Percentage of Hoogsteen Strand. *Data Quality Assurance and Region Annotations.* Prior to statistical analysis, a comprehensive quality check was performed on the data to ensure the integrity and accuracy of the sequence annotations. This included

https://doi.org/10.1021/acs.biochem.4c00334
Biochemistry 2025, 64, 250−265

verification of region-specific annotations such as 5′ UTR, CDS, and 3′ UTR segments. Additionally, sequences containing ambiguous nucleotides were excluded from subsequent analyses to maintain the robustness of our findings. For Figure S7, the following was followed: To comprehend the spatial distribution of triplexes within transcripts, the triplexes were classified into distinct categories based on their location. Using the minimum and maximum indices of each triplex, their position was ascertained within respective transcripts. If the start and end indices lay in the same region, the triplex was classified as belonging to lie in that respective region. If, however, the start index fell in the 5′ UTR and the end index in the CDS, the triplex was classified as a 5′ UTR-CDS spanning triplex, and if the start index was in the CDS and the end index in the 3′ UTR, this was classified as a CDS-3′ UTR spanning triplex. The triplexes were, therefore, delineated into five categories: 5_prime_utr_triplex (for triplexes in the 5′ UTR), CDS triplex (for triplexes in the CDS region), 3_prime_utr_triplex (for triplexes in the 3′ UTR), 5_prime_utr_CDS_triplex (for triplexes spanning the 5′ UTR and the CDS), and CDS_3_prime_utr_triplex (for triplexes spanning the CDS and the 3′ UTR).

Statistical Analysis. Statistical analyses were conducted using Python, leveraging robust libraries such as Pandas for data handling, SciPy for statistical testing, NumPy for numerical operations, and Matplotlib and Seaborn for data visualization. All of the data and statistical analyses were conducted in a Jupyter Notebook environment, which provided an interactive platform for data manipulation, analysis, and visualization.

The analysis pipeline included:

1. **Shapiro−Wilk Test**: This was employed to assess the normality of the GC content distributions across different transcript regions. This test was crucial for determining the appropriate statistical methods for group comparisons.
2. **Mann−Whitney U Test**: This test was used to compare the GC content between the full transcriptome and transcripts with triplexes, given the non-normal distributions observed. This nonparametric test is well-suited for analyzing median differences between two groups. The UTR regions showed particularly significant differences in GC content, with p-values less than 0.01, indicating a potentially functional role of varied GC content in these segments.
3. **Permutation Test**: The permutation test was conducted to further validate the significance of our findings and to account for variations in sample size. This test involved randomly shuffling data labels and recalculating mean differences thousands of times to estimate the probability of observing differences as extreme as those noted. The robustness of our findings is confirmed by permutation tests, with observed differences yielding p-values less than 0.001, thus confirming that these differences are statistically significant and not due to random variation.

Procurement of Oligos. The lyophilized and HPLC-purified sequences of unmodified RNA oligonucleotides were obtained from Genscript and dissolved in nuclease-free water. The solution concentrations of each of the oligonucleotides were determined optically at 260 nm and 25 °C by using the provided molar extinction coefficients (M^{-1} cm^{-1}) of strands. The folded triplex was obtained (as described earlier)[41] by heating the solutions in 10 mM sodium cacodylate buffer (with 150 mM NaCl and 0.5 mM $MgCl_2$) to 100 °C for 5 min and then keeping them on ice for 5 min, followed by a slow return to room temperature. All biophysical experiments were performed in a 10 mM sodium cacodylate buffer (pH 7.0) containing 150 mM NaCl and 0.5 mM $MgCl_2$ at 25 °C, unless otherwise specified. All experiments were replicated three times. The sequences used are shown in Table 1.

Thermal Difference Spectra Calculation. 1 μM folded RNA was taken in a quartz cuvette of 1 cm path length in the thermoelectrically controlled Cary 3500 (Varian) spectrophotometer and scanned from 350 to 200 nm, first at 15 °C, then at 95 °C. The absorbance values at 15 °C were subtracted from the absorbance values at 95 °C to give a plot of thermal difference spectra (of delta absorbance vs wavelength). The buffer used for TDS experiments was 10 mM sodium cacodylate (pH 7), 0.5 mM $MgCl_2$, and 150 mM NaCl.

CD Spectroscopy Analysis. CD spectra were recorded in a JASCO 815 spectropolarimeter equipped with a thermoelectrically controlled cell holder and a cuvette with a path length of 1 cm. CD spectra for the triplexes (at 2 μM) were recorded between 220 and 350 nm at 100 nm/min scan rate at 25 °C, and the spectrum obtained was the average of three scans. The buffer used for CD experiments was 10 mM sodium cacodylate (pH 7), 0.5 mM $MgCl_2$, and 150 mM NaCl.

UV Melting Spectroscopy and Derivative Calculation. Absorbance vs temperature profiles (melting curves) for triplexes (1 μM) were measured at 260 and 295 nm with a thermoelectrically controlled Cary 3500 (Varian) spectrophotometer. A temperature range of 15−95 °C was used to monitor the absorbance at the specified wavelength with a heating/cooling rate of 0.2 °C min^{-1}. First derivatives of the melting profiles were then computed using the formula $(y_2 - y_1)/(x_2 - x_1)$ for each data point to provide the y-axis values. The buffer used for UV melting experiments was 10 mM sodium cacodylate (pH 7), 0.5 mM $MgCl_2$, and 150 mM NaCl.

RESULTS

Development of a Bioinformatic Python- Based Code to Search for *MALAT1*-like Triple Helices. We targeted our search to the human transcriptome, hence characteristic features of a *MALAT1* or *NEAT1*-like triple helix (of human origin) were analyzed as a reference model (Figure 1a). An important feature present in both *MALAT1* and *NEAT1* triple helices and which has been accounted for in our script is the presence of a gap nucleotide in the Hoogsteen strand of the triplex. However, it has been shown by a more recent structural report that a triple helix with 11 consecutive base triples can also form a structure.[42] (This type of triple helix has not yet been found in nature, though.) For this reason, both possibilities were taken into consideration, and the search was carried out in two modes: one to search for a triplex with gap and another to search for a triplex without gap. We started our search with finding a fixed sequence length of Hoogsteen (H) strand, followed by looking for a downstream Crick strand that is similar to (in case of H sequence with gap) or identical to (in case of H sequence without gap) the Hoogsteen strand. Next, we looked for the presence of a further downstream Watson strand with a sequence complementary to that of the Crick strand (and also to that of the Hoogsteen strand). Toward this, the search criteria for length of Hoogsteen strand were fixed between 7 and 15 nt (for both cases—with and without gap). A sequence of length between 3 and 100 nt was accepted between the Hoogsteen and Crick strands, and a similar 3−100 nt sequence length was considered between Crick and Watson strands. Using the

https://doi.org/10.1021/acs.biochem.4c00334
Biochemistry 2025, 64, 250−265

Table 1. Oligonucleotide Sequences

Name of lncRNA	Sequence
MALAT1 94mer	GGAAGGUUUUUCUUUUCCUGAGAAAACAACACGUAUUGUUUUUCUCAGGUUUUUGCUUUUUGGCCUUUUUCUAGCUUAAAAAAAAAAAAGCAAAA
JPX	GAACUUUUUUUUUUUUAUCCAUUUGAAAGGUAUUUUUUUUUGGGUGAACUAUUUGUUAAUAUUCUUGGUCCUGUAAAAAAA
FTX	AAGAAUAUUUUAUAAGAACUCAUCUUAAUUUUUAUAAUUAUUAACCUCUUAAAUCUGUGUAAGACAUUCUAUUUAAUGCAACUCCUUAUAAAA
Xact	UGUUUGUCUUUUUUAAAAUAAAUUUUUACUGUGCAUAUUUAAGGUAUACAAUAUGAUGUUAUGAAACAUAUGUGUACAGUAAAA
lnc-BHLHB9-1-11	UCAUAUAAUUUAAUACUAUAAUCUAAUUUAAUAUUACAAUUUUAUAAUGUAAUUUUAAUAAUAUUAUAUUAAAUUAAAAUUA
lnc CHMP2B-5-1	UGUUAUCAAUAAUAUUUGGUUGAUCUUCACAUAUUUUAUAUGCAUAUAUAUCAAGAAGUUAUAUAUAUAACUCAAGAAACUCAAGUUAUAUAUA
lncMSR107:1	ACAUACAUAUAAAAUUAUAUUUUUAUGUAAAAUUAUAUAUAUAAAAAUUGACAGUAAUUUUUAUAUAU
lncKBTBD3-3-1	UUGGUAUUAAUUAUUAAUUAACCUAUUAAAUUAUUAAAUUAUAAUUUAAUA
Novel (ENST00000624837.1)	AUAAUAUAAUAUAUAAUAUAGAGUACAUAUUUAUAUUGUAUAUAAAUAUAUAAACAAUAUGAAUAUAUAUUUAUAU

mentioned constraints, the logic of the code returned Triplex if (and only if) it finds the Hoogsteen−Crick−Watson triplex followed by upper stem or upper loop and lower stem or lower loop. In all other cases, a triple helix was not returned. As is clear from these criteria, any kind of ribonucleotide base permission constraints were excluded. This allowed any sequence of ribonucleotides to occur in a strand, be it a homopurine stretch, homopyrimidine stretch, or even a heteropolymer containing a mixture of purines and pyrimidines. (Filtering for presence of polypyrimidines has been done in a subsequent step, and findings were reported in different categories.) Further, the whole length of the sequence under consideration was searched using this algorithm. We are cognizant that if the triplex that is found is not toward the 3′ end of a transcript, it would be unlikely to perform the role of a stability element. However, this was done by design to query and explore new locations where hits are found. It was hypothesized that an RNA triple helix structure may also have functional roles such as protein recruitment and binding, other than their canonical properties of stabilizing the transcripts and increasing their half-life, and that is the basis of this broad criteria. With regard to the stem loop, the 3−100 nt gap sequences at both ends of the triplex were searched for to contain complementary base pairing, after the Hoogsteen−Watson−Crick (H−W−C triplex) strands are successfully found. The algorithm searched for an upper stem base pairing in the sequence between the end index of the H Strand and the start index of the C strand. If a stem is found, the remaining sequence is reported as the upper loop of the stem; otherwise, the entire segment is returned as an upper loop only. In both cases, the upper stem and the upper loop have a length constraint of a minimum of 3 nt for each. Similarly, the algorithm searched for the formation of a lower stem loop by complementary base pairing between the gap ribonucleotides in the C−W strand and the sequence preceding the H strand using the same constraints of a minimum length of 3 nt. The presence of stem in the locations akin to the P1 and P2 regions of MALAT1 triple helix is not a mandatory criterion; the positions and sequences are reported if found, otherwise considered a loop. This complete tool was named TRIPinRNA (Figure 1b,c). Two versions of it have been created—one which searches for a gap nucleotide mandatorily in the Hoogsteen strand (with-gap mode) and one version which searches for a continuous Hoogsteen strand (called the “without-gap” mode). In the case of search “with-gap”, the gap may occur at any one position in the Hoogsteen strand, with the exception of three nucleotides at the 5′ and 3′ extremes of the Hoogsteen strand.

The search for intramolecular RNA triplexes was made in the partial NONCODE database, complete GENCODE lncRNA database and also the GENCODE coding data set.[1,40,43] A sample output is shown in Figure S1. The complete output files from the runs are shared in excel format in Files S1−S4.

TRIPinRNA Predicts Triple Helix Formation in Human lncRNA Sequences. TRIPinRNA was run on the GENCODE lncRNA data set, which contained a total of 58,246 transcripts (release 44 for human—GRCH38.p14), and revealed 3178 predicted triplex hits for TRIPinRNA run in “with-gap” mode and 10,769 predicted triplex hits for TRIPinRNA the run in “without-gap” mode. Out of these, the number of unique transcripts was 2788 for the run in “with-gap” mode, which was mapped to 1509 unique genes when considering run in “with-gap” mode. In the case of the run in “without-gap” mode, the number of unique transcripts was 7936, and the number of unique genes that these mapped to was 3762 (Table 2).

https://doi.org/10.1021/acs.biochem.4c00334
Biochemistry 2025, 64, 250−265

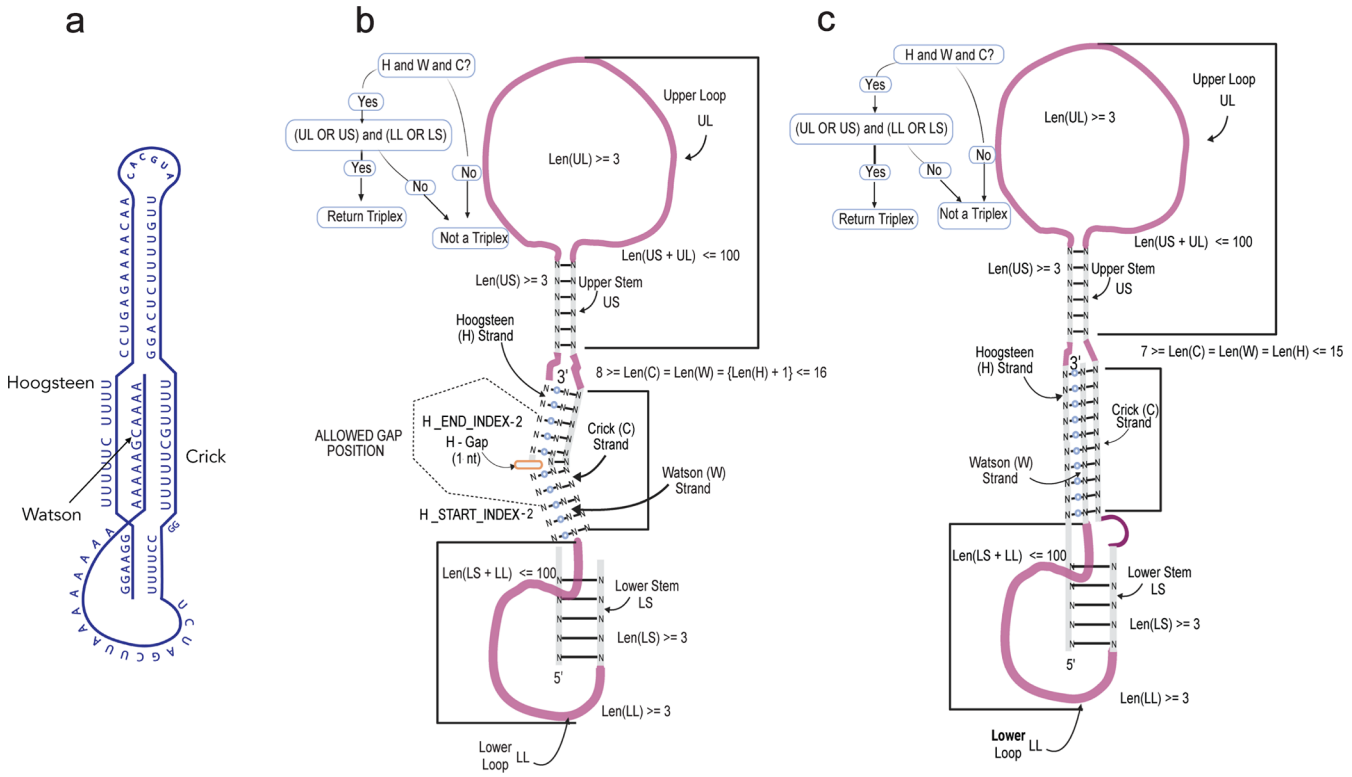


Figure 1. RNA intramolecular triple helices with Hoogsteen (H), Watson (W), and Crick (C) strands labeled (a) 94mer MALAT1 triple helix. Schematic for TRIPinRNA search algorithm in (b) "with-gap" mode and (c) "without-gap" mode.

Table 2. Summary and Classification of Predicted Hit Triplexes from the Long Noncoding Human Transcriptome from the GENCODE Database[a]

	lncRNA (Noncoding)							
	With Gap				Without Gap			
	Total	Pyrimidine- Only	Uridine-Only	Cytidine-Only	Total	Pyrimidine-Only	Uridine-Only	Cytidine-Only
No. of Triplexes	3718	234	16	1	10,769	794	77	0
No. of Unique Transcripts	2788	277	16	1	7936	735	77	0
No. of Unique Genes	1509	131	15	1	3762	396	34	0

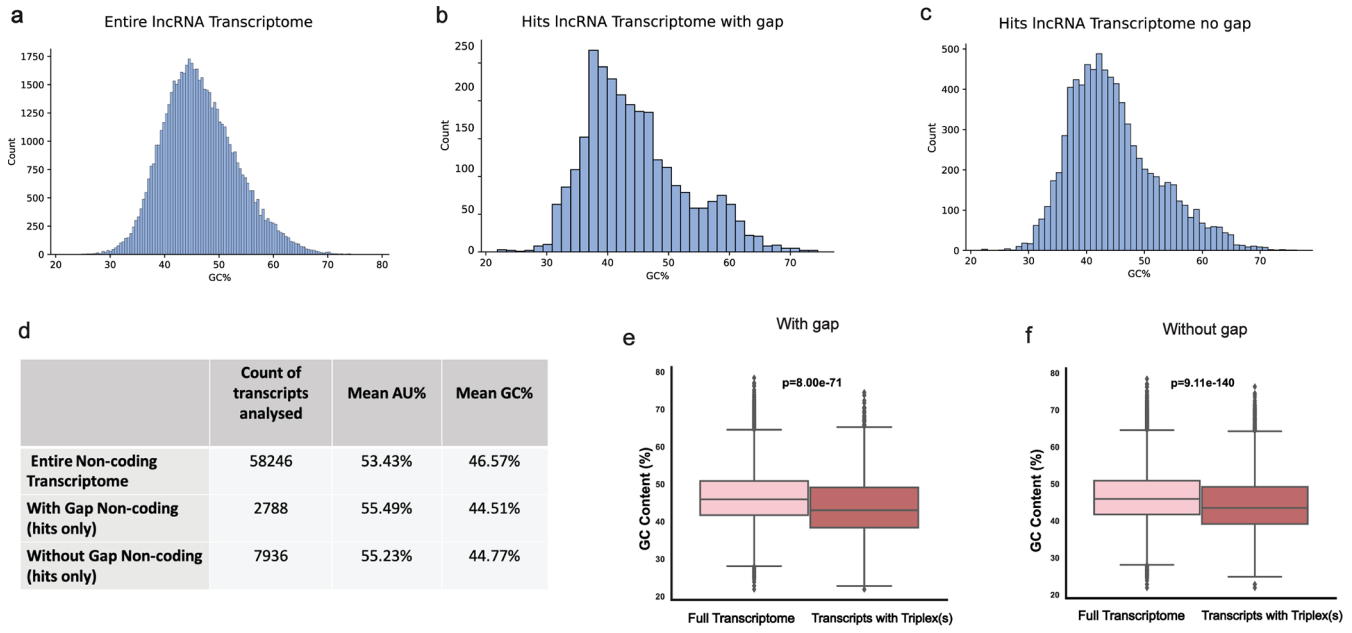
[a]Pyrimidine -only—triplex hits with only U or C in the Hoogsteen strand, uridine- only—triplex hits with only U in the Hoogsteen strand, cytidine-only—triplex hits with only C in the Hoogsteen strand.

The distribution of GC percent usage was further plotted for the entire human noncoding transcriptome (Figure 2a) and compared to that of only the transcripts containing hits (Figure 2b,c). It was found that there was a small but significant decrease in the mean GC percent from 46.57% to 44.51% in the case of run in "with-gap" mode and 44.77% in the case of run in "without-gap" mode (Figure 2d–f). Significance calculations have been shown in Table S1.

To elucidate any inherent preferences in the lengths of the Hoogsteen strands, frequency distribution plots were generated. These plots were prepared for three distinct scenarios: for overall triplex hits (these hits included the Hoogsteen strand of the triplex portion containing a combination of purines and pyrimidines), for those hits with only pyrimidine-based Hoogsteen strands (U and C), and finally, for hits with only Uridine in the Hoogsteen strand. In the plot for overall triplex hits, a gradual decrease was observed in the number of predicted triplex hits, with increasing length of the triplex from 7 to 15 nt in the Hoogsteen (H) strand. This was accompanied by a slight increase in length 13 nt for the "with-gap" mode and a slight increase in length 15 nt for "without-gap" mode (Figure 3a,b). Filtering the predicted triplex hits for "pyrimidine-only" (to represent the triplex nucleotide composition expected in a canonical pyrimidine-type triplex) gave 234 triplexes as hits in the "with-gap" mode and 794 triplexes in "without-gap" mode. Only triplex lengths of 7, 8, 10, and 13 nt were observed in "with-gap" mode, and triplex lengths of 7, 8, 9, and 13 nt were observed for "without-gap" mode (Figure 3c,d). Filtering for "uridine-only" in the Hoogsteen strand yielded 16 triplexes in the "with-gap" mode and 77 triplexes in the "without-gap" mode. With this criterion, there were triplexes with lengths of 7, 8, and 13 nt only in the Hoogsteen strand for the "without-gap" mode and triplexes of lengths 7, 8, and 13 nt only for "with-gap" mode (Figure 3e,f).

Next, we asked how many transcripts had multiple triplexes predicted in a single transcript. This was also visualized by using a plot. It was observed that most of the transcripts contained up to 5 triplexes for run in "with-gap" mode and up to 11 triplexes for run in "without-gap" mode. Nonetheless, there were a few transcripts with a greater number of triplexes. For example, in the "all, with-gap" mode, transcripts with 39 triplexes (XACT), 15 triplexes (HELLPAR lncRNA), 10 triplexes

https://doi.org/10.1021/acs.biochem.4c00334
Biochemistry 2025, 64, 250−265



	Count of transcripts analysed	Mean AU%	Mean GC%
Entire Non-coding Transcriptome	58246	53.43%	46.57%
With Gap Non-coding (hits only)	2788	55.49%	44.51%
Without Gap Non-coding (hits only)	7936	55.23%	44.77%

Figure 2. Overview of GENCODE lncRNA data set results. Histogram of GC percent distribution calculated in transcripts: (a) from the entire human noncoding transcriptome, (b) with predicted triplex hits after TRIPinRNA run in "with-gap" mode, and (c) with predicted triplex hits after TRIPinRNA run in "without-gap" mode. (d) Tabulated number of transcripts analyzed along with their GC percentage. Difference in GC usage plotted with significance values noted: (e) for run in "with-gap" mode and (f) for run in "without-gap" mode.

(ENST00000620188.1), and 7 triplexes (MIR2052HG) were observed (Figure 4a), and in "without-gap" mode, transcripts with 97 triplexes (XACT lncRNA), 41 triplexes (HELLPAR lncRNA), 27 triplexes (ENST00000661959.1), and 10 triplexes (ENST00000570269.2) were observed (Figure 4b). In the "pyrimidine-only, with-gap" mode, the maximum number of triplexes is two, which is present in 7 transcripts, and in the "pyrimidine-only, without-gap" mode, the maximum number of triplexes is four, which is present in a single transcript (Figure 4c,d). With the "uridine-only" filter on hits, all transcripts in "with-gap" and "without-gap" mode have only 1 triple helix predicted per transcript (Figure 4e,f), An overlap was noticed between gene names with multiple triplexes in "with-gap" and "without-gap" mode, and the intersection between these in "uridine-only" type canonical triplexes was determined. Seven predicted genes were found, one of which was *JPX* (Just Proximal to Xist), an lncRNA with a role in X Chromosome Inactivation.[44] The predicted RNA triplex of this gene was picked for further study in detail (Table 3).

Biophysical Validation of Triple Helix Sequences Predicted Using TRIPinRNA. To validate predicted hits, we have utilized circular dichroism (CD) spectroscopy and ultraviolet (UV) melting. Spectroscopy methods are widely used to investigate RNA structures and can be utilized to study the formation of a triple helix motif too. The predicted 80-nucleotide region of *JPX* was chemically synthesized (Genscript, HPLC purification) (Figure 5a) and used for validation. An appropriate amount of the oligo was heated and subjected to slow cooling to preform structure in sodium cacodylate buffer (containing NaCl and $MgCl_2$), then used for biophysics experiments. On thermal melting for the predicted *JPX* triple helix sequence, a relatively stable triplex structure with two melting temperatures was observed—one at around 58 °C and one around 75 °C (Figure 5b). This is similar to the melting pattern (with two transitions) as has been seen earlier for the *MALAT1* triple helix motif (Figure 5c). Next, CD spectroscopy showed that the signature for the *JPX* sequence was practically identical to that seen for the *MALAT1* 94mer sequence, with two positive (at 260 and 220 nm) and two negative peaks (at 240 and 210 nm) (Figure 5d,e). A similar pattern in CD signature has been reported previously for the KSHV PAN triple helix, too[45] and also for a model UAU type intramolecular triple helix,[46] which is an important comment on the structure formed in predicted *JPX* triple helix sequence.

A few other genes came up as hits containing triple helix sequences which were involved in X chromosome inactivation (XCI) under the "all" list. Including *JPX*, this was a total of 6 genes, which was intriguing. Therefore, these five other XCI genes (*FIRRE*, *FTX*, Xist, Tsix, and Xact)[47−52] that had shown to have predicted triple helix sequences were taken up for investigation (Table 4). These were, however, not the canonical poly-U, poly-A, and poly-U strands containing triplexes (Figure 6a). These were hits that had been obtained from the "all" search and had strands that could and did contain a mixture of purine and pyrimidine ribonucleotides. Further, the predicted 3′ most triple helices in these genes were not located near the 3′ end of that gene, as may be expected if they were to act as stability-element type triplexes. Nonetheless, these were of great interest to us. For example, the transcript of Xact—which is 347,561 nt in length—has 97 triple helices predicted to form in it in the "without-gap" mode and 39 triple helices predicted to form in it in the "with-gap" mode, and the locations of these triple helices are well-distributed across the transcript (Figure S2). Keeping these points in mind, we moved ahead to investigate these predicted sequences for triple helix structure formation as seen by biophysics. Out of the five genes containing predicted nonpoly pyrimidine-type triplexes, two hits were chosen (within 100 nt in length) for which the RNA was synthesized. These are (1) the predicted 92 nt triplex helix closest to the 3′ of *FTX* (five prime to Xist) and (2) the predicted 83 nt triple helix closest to the 3′ of *XACT* (X active specific transcript). The UV melting signature of *FTX* showed a melting temperature of around 53 °C

https://doi.org/10.1021/acs.biochem.4c00334
Biochemistry 2025, 64, 250−265

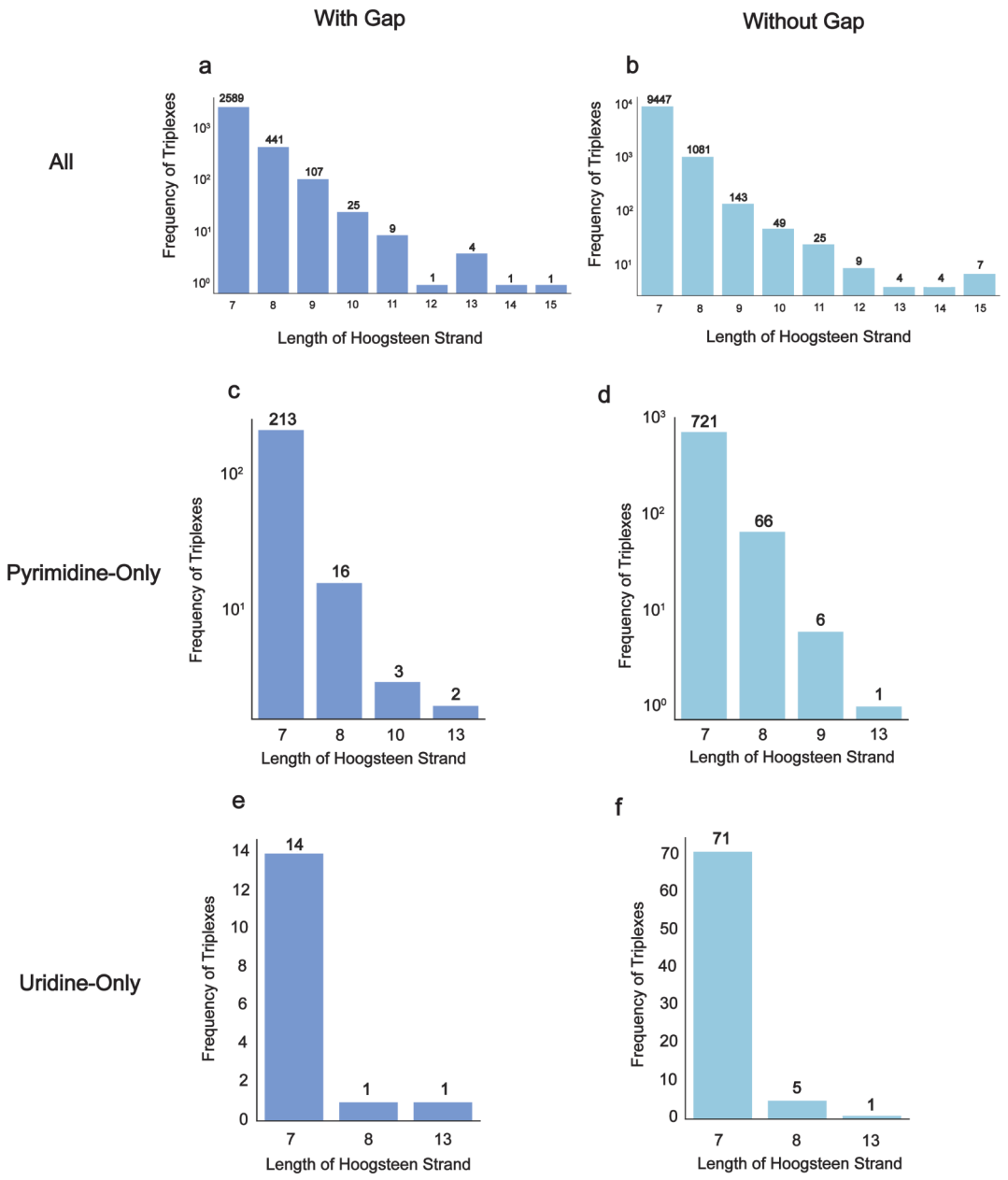


Figure 3. Distribution of triplex lengths (length of Hoogsteen strand) for predicted triplex hits from TRIPinRNA run on the GENCODE human lncRNA data set: (a) with "all" output filter obtained in "with-gap" mode, (b) with "all" output filter obtained in "without-gap" mode, (c) with "pyrimidine-only" output filter in "with-gap" mode, (d) with "pyrimidine-only" output filter in "without-gap" mode, (e) with "uridine-only" output filter in "with-gap" mode, and (f) with "uridine-only" output filter in "without-gap" mode.

with a rather broad melting. (Such a broad melting domain is generally observed when the melting temperatures of two or more domains are close together and overlap such that discrete melt peaks are not observable.) The melting profile of *XACT*, however, showed three melting domains, with melting temperatures of around 30, 45, and 60 °C. This correlated well with the predicted structure of *XACT* as observed from UNA RNA Fold,[53] which predicted a structure forming in the sequence of the bulge region of *XACT*, in addition to the expected two melting domains (of unraveling of the Hoogsteen strand and then the duplex unwinding). Further, the CD signatures for both triplexes were practically identical to what was observed in the case of *MALAT1* 94mer, with two negative and two positive peaks, indicating the formation of an intramolecular RNA triple helix (Figure 6b). Thermal difference spectra (TDS) is a characteristic graph pattern obtained by subtracting the absorbance scan at high temperatures with the respective scan at a low temperature and is used to identify the structure formed by the sequence.[54] The TDS for these two triplexes also showed a typical signature with a positive peak at 260 nm and a continuous shoulder at 245 nm. The TDS additionally had a negative peak at 295 nm, indicating the presence of a Hoogsteen bond in the structure (Figure 6b). All these signatures were representative of an RNA triple helix structure and were very similar to the respective signatures for *MALAT1's* 94mer and demonstrate that these noncanonical type of triple helices are also capable of forming structure. The thermal melting at 295 nm for these RNA structures has also been performed, and the temperature where the dip in absorbance is observed coincides with the breaking of the Hoogsteen bonds between the H and W

https://doi.org/10.1021/acs.biochem.4c00334
Biochemistry 2025, 64, 250−265

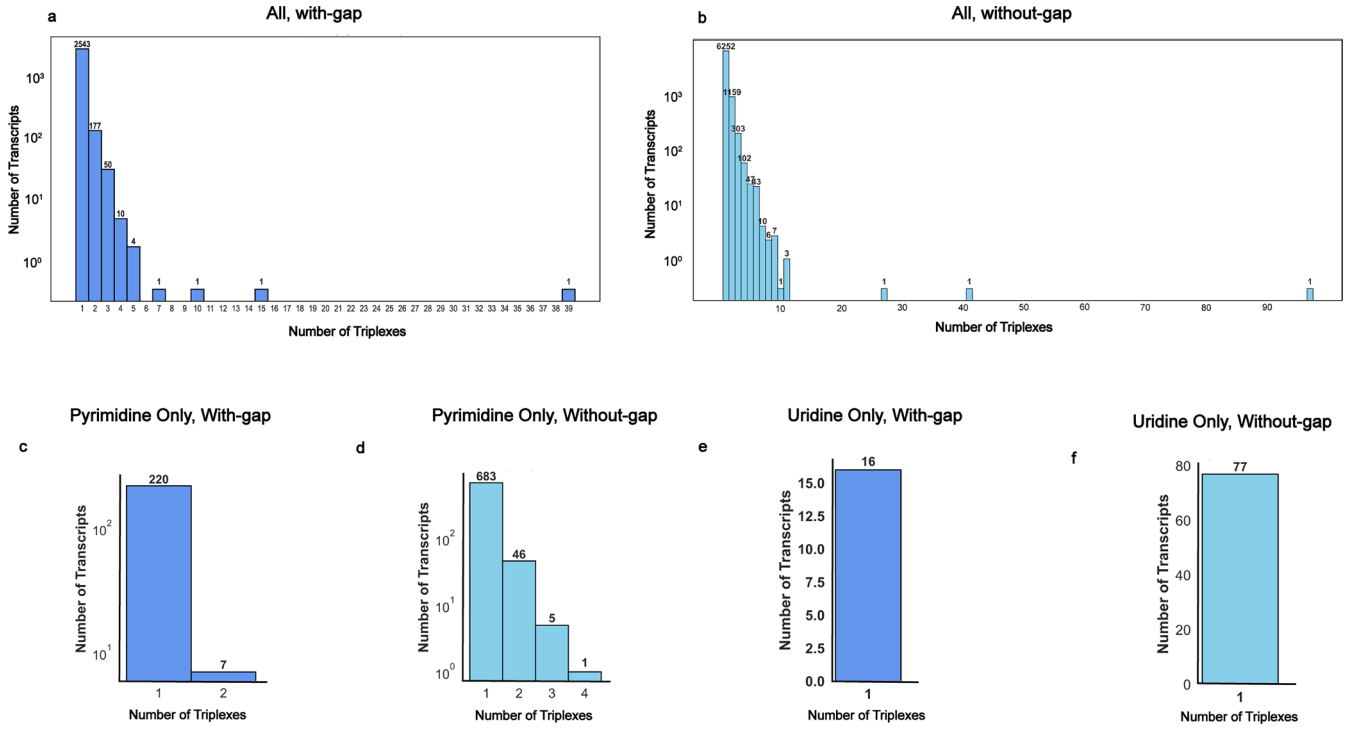


Figure 4. Frequency distribution of transcript hits with multiple triplexes obtained from TRIPinRNA run on the GENCODE human lncRNA data set (a) in "all" output filter "with-gap" mode, (b) in "all" output filter "without-gap" mode, (c) in "pyrimidine-only" output filter "with-gap" mode, (d) in "pyrimidine-only" output filter "without-gap" mode, (e) in "uridine-only" output filter "with-gap" mode, (f) in "uridine-only" output filter "without-gap" mode.

Table 3. List of Ensembl Gene Names and Descriptions of Common Transcripts between "With-Gap" and "Without-Gap" Type Searches for Poly-Uridine Triplexes Predicted

	Gene name	Description
1	ENSG0000000267519.6	miR-23*a*/27a/24-2 cluster host gene
2	ENSG0000000258634.3	Novel, antisense to ALX3
3	ENSG0000000279512.1	Novel, uncategorized
4	ENSG0000000286125.3	(Novel Transcript (Contains ZIM2-AS1 and PEG3-AS1))—Uncategorized gene
5	ENSG0000000225470.9	JPX transcript XIST activator
6	ENSG0000000260206.1	ENSG0000000260206 (Novel Transcript, lncRNA- Antisense To IMP3)
7	ENSG0000000234208.1	ENSG0000000234208 (Novel Transcript, lncRNA-Sense Intronic To THOC5)

strands (Figure S8). This structure, however, is less thermally stable than that seen with the canonical pyrimidine-type helices. A modified triple helix control sequence in contrast, as shown in an earlier report, shows a much different pattern of UV melting as well as for CD spectroscopy.[41]

This observation prompted us to query other noncanonical triplexes that turned up as hits from a run on the NONCODE data set of noncoding RNAs.[40] A set of 5 potential triple helix-forming RNAs was selected, for which RNA was prepared using synthesis (Genscript, HPLC purification). These five are the following: lnc-BHLHB9-1-11, lnc CHMP2B-5-1, lncMSR1-7:1, lncKBTBD3-3-1, an RNA which was termed "Novel" by us and which belongs to the transcript ENST00000624837.1 (Figure S3). The synthesized RNAs were pre formed for structure as described in methods. The UV melting, CD, and TDS signatures for these sequences were indicative of the formation of a triple helix structure (Figure 7a,c). In general, these type of triple helices have not been reported before to form and are speculated to be unstable as it is believed that any base triplet other than a UAU or CGC (forming a pyrimidine-type triplex) is destabilizing.[16] However, our observations indicated otherwise, albeit at the biophysics level (in vitro).

Predicted Triplex-Forming Sequences Observed in the Coding Region of the Genome. Next, we asked, what is the situation with predicted triplex-forming sequences in the protein-coding human transcriptome? Is it possible to observe any differences in the type or prevalence of predicted triple helix-forming sequences between the noncoding and coding regions? Our guess was that noncoding regions would have more use for a structure-forming region than any coding region. However, on running TRIPinRNA on the set of coding sequences (110,962 in number) downloaded from the same release of GENCODE (release 44 for human—GRCH38.p14), a much higher number of predicted triplex-forming sites was found than in the long noncoding transcriptome. Overall, in the run "with-gap" mode, there were a total of 14,561 triplexes predicted, and in the run "without-gap" mode, there was an extraordinary number of 90,836 triplexes predicted. Out of these, only 4 triplexes from the run in "with-gap" mode and 8 triplexes from the run in "without-gap" mode were CGC type (Table 5).

The histogram for GC distribution across the entire protein-coding transcriptome showed a small spike around 40% and another at around 60% (Figure 8a). When considering the GC distribution for only transcripts containing hits, a bias was observed (Figure 8b,c). It was seen that there was a very significant decrease in the case of 3′ UTR region as well as a very significant increase in the case of the 5′ UTR region in both run modes (Figure 8d). As for the CDS, there was a very small and insignificant change in the case of run in "with-gap" mode. For run in "without-gap" mode, a 0.01% decrease in average GC usage was observed, with a p value of 0.006 (Figure 8e,f). From

https://doi.org/10.1021/acs.biochem.4c00334
Biochemistry 2025, 64, 250−265

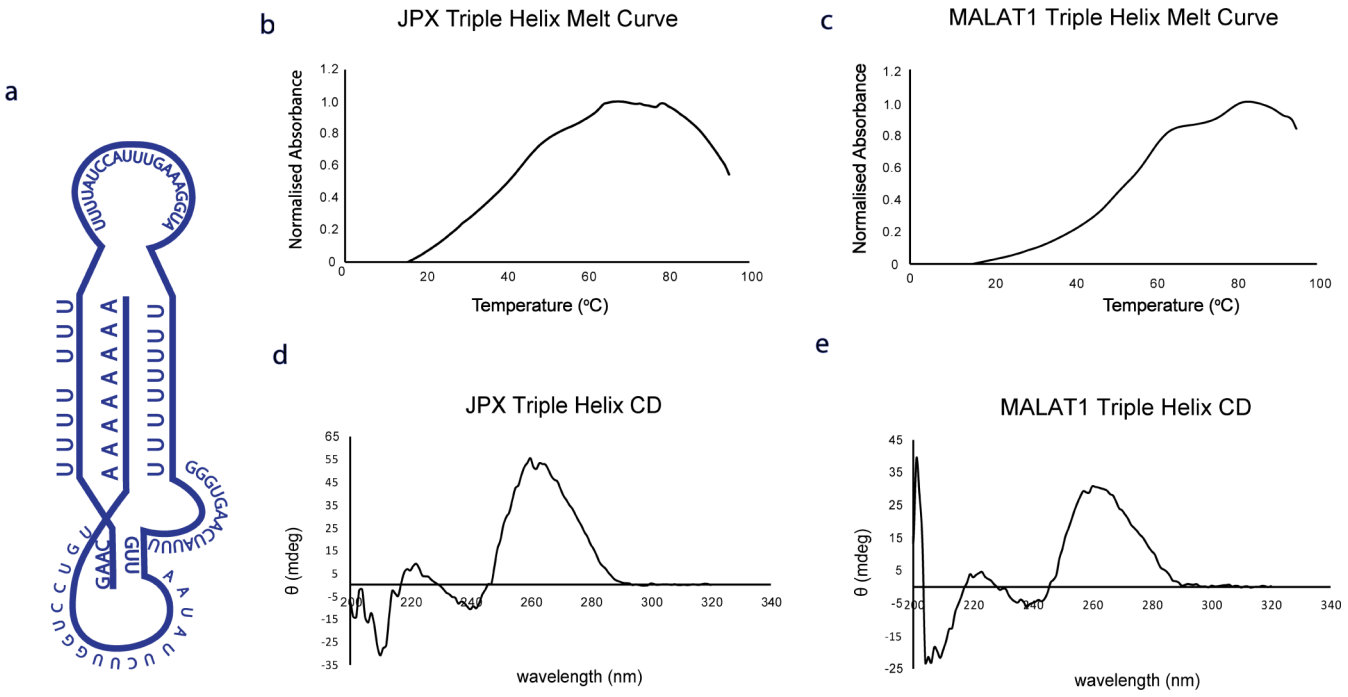


Figure 5. Canonical UAU-type *JPX* triple helix and its biophysical characterization compared with that of MALAT1 94mer triple helix: (a) predicted *JPX* triple helix structure, (b) thermal melting for *JPX* triplex, (c) thermal melting for 94-mer triplex from MALAT1, (d) CD spectra for JPX triplex, and (e) CD spectra for 94mer triplex from MALAT1.

Table 4. Summary of XCI-Related Genes with Predicted Canonical and Noncanonical Triplexes

Name	Number of Transcripts with Triplexes	Number of Unique Triplexes
Xist	11	1
Tsix	1	1
FTX	8	4
JPX	1	1
FIRRE	1	1
Xact	1	39

this observation, it was inferred that although triplexes are predicted across the entire coding transcript, there seems to be a preference by the UTR regions toward a certain altered GC usage, while this preference is absent from the CDS regions. We speculate that since the mutational rate is higher in the UTRs as compared to the CDS region, the evolution of triplex-forming sequences in the UTR regions is more visible in the UTRs.[55,56] Further, this is in congruence with the presence of other RNA structural elements in UTRs carrying out different regulatory functions.[57,58] The significance values in support of these statistical analyses have been shown (Table S1).

To further delve into the nucleotide composition of the Hoogsteen strands of hits, a thorough investigation was conducted into their GC and AU content. For triplexes with a gap in the H strand, the gap, represented by a ".", was excluded from our calculations to ensure accuracy. (In noncoding hits, it was seen that the triplexes tend toward using a high AU percentage on average in both modes of run.) (Figure S4) In the coding region, while it was observed that the maximum number of predicted triplex hits lay in the 3′ UTR (Figure 9a,b), there was more variation with regard to GC usage in triplex hits (Figure S5a). The hits of the CDS portion exhibit high GC usage, on average (Figure S5b). The hits of the 5′ UTR again exhibit a high average GC usage (with a mean of 84%) in their H strand in both modes of run ("with-gap" and "without-gap") (Figure S5c). Lastly, the 3′ UTR hits exhibit a very low GC usage (of 22%) in their H strand on average in both modes of TRIPinRNA run (Figure S5d). It is relevant to note that a low to average correlation was observed when the number of triplexes per transcript was plotted against the length of transcripts, indicating that a longer transcript does not guarantee a larger number of hits. Hence, these hits are not a result of chance, but rather, are functionally relevant (Figure S6). Overall, the hits lying in the 5′ UTR were the highest in GC usage on average, followed by the CDS hits, and the hits of the 3′ UTR had the lowest GC usage, on average (Figure S7a,b). Lastly, a crude calculation was performed to figure out the density of predicted triplex hits in different regions of the transcriptome. This was divided into four categories: (1) noncoding RNAs, (2) 5′ UTR, (3) CDS, and (4) 3′ UTR. The number of hits in that category was divided by the cumulative length of that region in transcripts with the hit. For example, to calculate the density of the 5′ UTR, the number of triplex hits that fall in the 5′ UTR was counted and divided by the cumulative length of the 5′ UTRs of all transcripts that reported a hit in the 5′ UTR. This allowed us to understand that the highest density of hits for both "with-gap" mode and "without-gap" mode was in their 5′ UTR regions, and taken overall, the highest density was in the 5′ UTR in "without-gap" mode (Table S2). The CDS had the lowest number of predicted triple helix hits, in both modes of search. In this analysis, however, the triplexes that lie on the two boundary regions between the UTRs and the CDS have not been taken into account. (Those number of hits were <7% overall and have been ignored for the purpose of this analysis.)

In conclusion, we have seen that a few new canonical pyrimidine-type and uridine-only triple helices have been seen to form in some human genes. We also experimentally show triple helix structure formation in noncanonical (heteropurine−

https://doi.org/10.1021/acs.biochem.4c00334
Biochemistry 2025, 64, 250−265

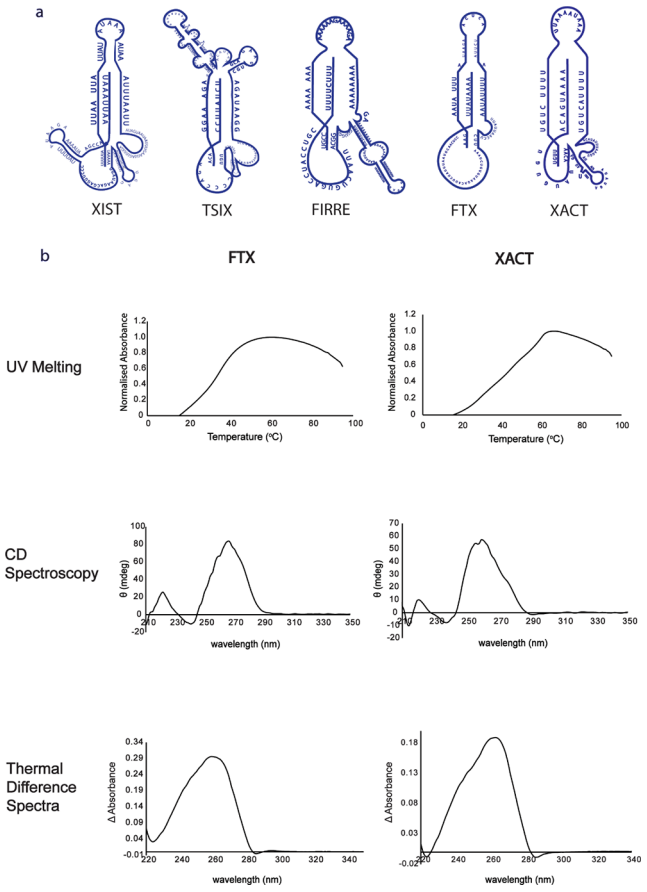


Figure 6. XCI transcripts predicted by TRIPinRNA to contain heteropurine heteropyrimidine strand-type triplex hits—Xist, Tsix, *FIRRE*, *FTX*, and Xact. (a) Possible folding patterns of the predicted sequences. (b) Thermal melting, TDS, and CD for two noncanonical XCI triplexes *FTX* and *XACT*.

pyrimidine type) triple helices. Further, there are triple helices predicted to form not only in the noncoding portion but also in the coding portion of the human genome. Interestingly, while the noncoding portion of the genome predominantly gave AU-type triple helices, the coding portion showed predicted hits that had GC usage, which was very dependent on the location in the transcript. With this large set of thousands of predicted triplexes, a comprehensive investigation into this structure is now possible, which could possibly aid functional studies. A systematic study could also correlate the characteristics of loop length, triplex strand length, presence of stem, and other such characteristics with the temporality and stability of triple helix structures. Also opened up by this body of work is the possibility of using this data set for machine learning to train and then predict newer locations for these triplexes. Moreover, this work could be extended to genomes of different organisms and also extended in different directions such as correlating the presence of predicted triplex sequences with phenomenon such as nonsense mediated decay (NMD) when these structures occur in the CDS, for example.

DISCUSSION

In this study, we explored the presence of potential intramolecular RNA triplex-forming sequences in the human genome. Taking a cue from the known intramolecular RNA triple helices (*MALAT1* and *NEAT1*), a python program was developed using a "sequence-pattern" search methodology based on the characteristic patterns observed in these two known triplexes. We started out with a broad criterion: searching for parallel-type triple helices without sequence constraints in the Hoogsteen strand. This allowed for results to include noncanonical heteropurine−pyrimidine type triplexes. Though it is an unconventional thought that such triplexes could form, we observed using various biophysics methods that these triplexes indeed do form in vitro. The major groove of a Watson−Crick double helix RNA, it is understood, can spatially not accommodate the binding of two purines in a triplet. For example, while the replacement of a C.G-C triplet of MALAT1 triple helix with an A.U-A triplet is well-studied to be destabilizing (by using the technique of the beta globin reporter assay), that such triplexes would never form with a temporality has not been negated.[16] Another report studies the A.U-A triplet stretch (to form a purine-type triple helix), but again, does not query the outcome in case of a heteropurine−pyrimidine strand.[59] However, recent reports have broadened our understanding of these complex structures. For example, in self-splicing introns and also in riboswitches, the types of triplets observed are often a noncanonical type of base pairing, indicating that the presence of these type of triplets is indeed possible.[60]

Additionally, the formation of such heteropurine−pyrimidine strand triplexes has been widely studied in the case of DNA triple helices. Howard et al. have characterized a novel Poly(dT)·2Poly(dA) triple helix, suggesting an expanded range of possible triplex configurations under specific ionic condition.[61] Further, Gondeaut et al.'s investigations into the interactions of methyl green and ethidium bromide with DNA triplexes shed light on the nature of these structures, potentially mirroring RNA triplex behaviors too.[62,63] The work of Lee et al. on 5-methylcytosine's influence in enhancing triplex stability at neutral pH further supports the possibility of the existence of these rather noncanonical triplexes.[64] Hoyne et al.'s identification of potential intrastrand triplex elements in bacterial genomes suggests a prevalence of these structures that might parallel those within the human transcriptome, and He et al.'s thermodynamic analysis provides crucial insights into the stability of DNA triplexes, informing our understanding of RNA triple helices.[35,65]

In view of the above understanding, we started our search with a broad criterion (without any sequence restriction in the first step of searching for the Hoogsteen strand) and followed up by filtering the output for results to also identify ones with only poly homopyrimidine Hoogsteen strands. This method also allowed us to account for a gap nucleotide around the middle of the Hoogsteen strand (as observed in the case of *MALAT1*), while allowing for a corresponding doublet that may break the poly homopyrimidine or poly homopurine criteria in the Watson and Crick strands (as seen in the case of *MALAT1* and *NEAT1* triplexes). Following this basis, we began our search.

The resulting search revealed that canonical (U.A-U type) as well as noncanonical (heteropurine−pyrimidine) type triplexes were predicted to exist across the human transcriptome. While the U.A-U type triplexes seemed to display greater thermal stability in vitro, the noncanonical triplexes showed typical signatures of triple helix structure formation using CD spectroscopy, UV melting, and thermal difference spectra. The UV melting profile showed either two or more (multiple) resolved peaks or a broad melting to indicate the sequential melting of different structured regions. CD spectroscopy

https://doi.org/10.1021/acs.biochem.4c00334
Biochemistry 2025, 64, 250−265

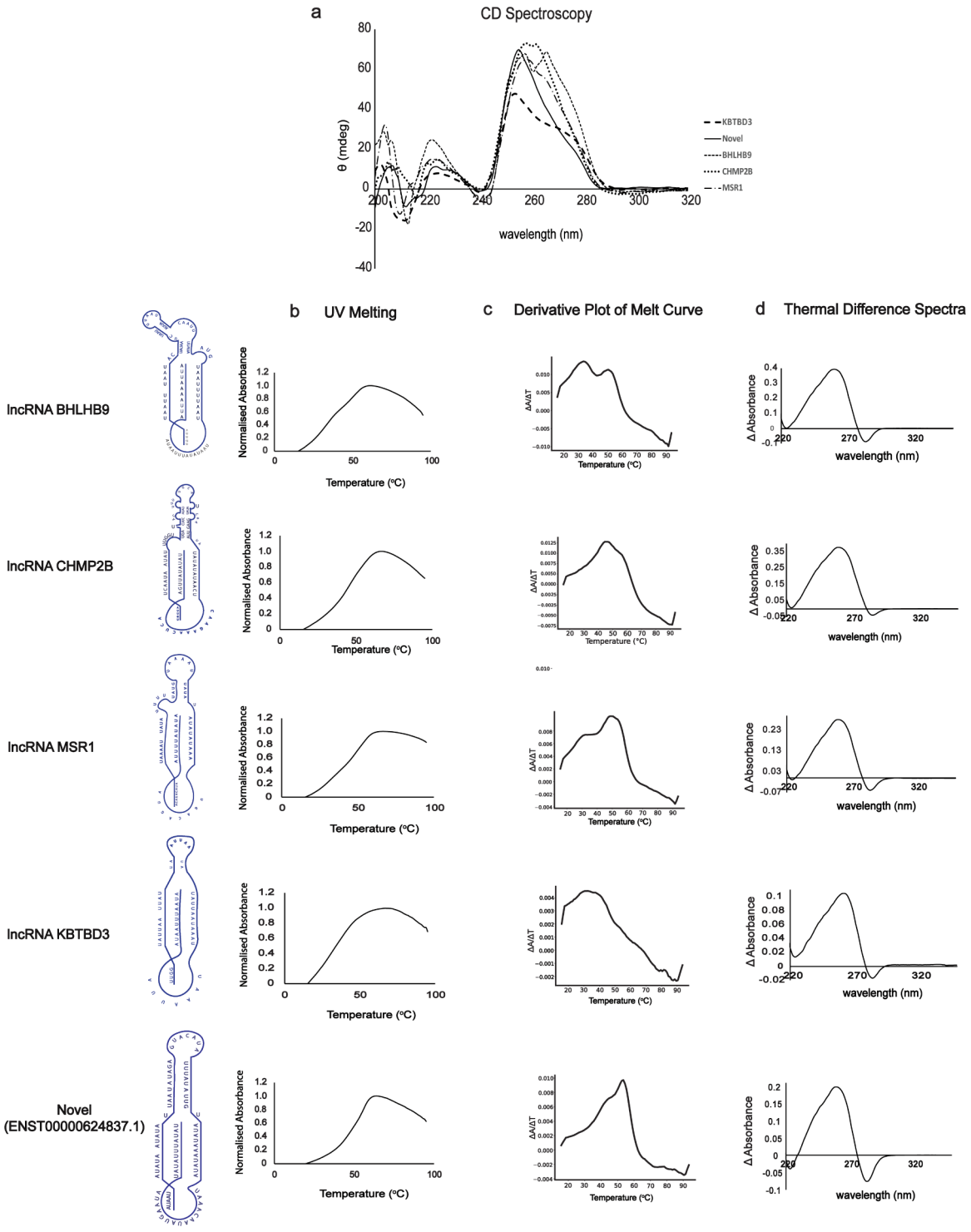


Figure 7. Biophysics for five heteropurine heteropyrimidine strand-type triplexes: (a) CD spectroscopy, (b) thermal melt curves, (c) thermal melting derivative curves, and (d) thermal difference spectra (TDS).

showed clear two positive and two negative peaks, and the amplitude of these peaks differed from that of A-form RNA. (In addition to the major positive peak at 260 nm, the positive peak at 220 nm was approximately equal in magnitude to the negative peak at 240 nm. Further, the negative peak at 210 nm was not as large in magnitude as that observed for A-form RNA.) Lastly, the thermal difference spectra pattern demonstrated a negative peak at 295 nm along with a positive peak at 260 nm. The TDS curve then almost touches the axis at 210 nm. This consistent TDS pattern might be the signature of an intramolecular triple helix. In the case of JPX and MALAT1 triple helix, the TDS pattern looked almost identical to this, except that a negative peak at 295 nm was not observed (data not shown). This also matches the report by Mergny and Chaires, where they do not observe a negative peak at 295 nm for only their T.A-T triple helices.[54] In summation, we recognized that these noncanonical triple helices could indeed form, and we continued to search in this mode—looking for "all" types of triple helices in the transcriptome.

https://doi.org/10.1021/acs.biochem.4c00334
Biochemistry 2025, 64, 250−265

Table 5. Summary and Classification of Predicted Triplex Hits from the Coding Portion of the Human Transcriptome[a]

	All[a]																			
	With Gap							Without Gap												
	Total	Pyr Only	U Only	C Only				Total	Pyr Only	U Only	C Only									
No. of Triplexes	14,561	963	197	4				45,418	2848	200	4									
No. of Unique Transcripts	11,315	845	190	4				27341	2537	192	4									
No. of Unique Genes	4556	357	82	4				9622	1017	93	3									
	Only CDS				3′ UTR							5′ UTR								
	With Gap			Without Gap				With Gap				Without Gap				With Gap			Without Gap			
	Total	Pyr Only	U Only	C Only	Total	Pyr Only	U Only	C Only	Total	Pyr Only	U Only	C Only	Total	Pyr Only	U Only	C Only	Total	Pyr Only	U Only	C Only	Total	Pyr Only	U Only	C Only
No. of Triplexes	3866	128	5	0	13,400	549	0	0	7578	657	175	2	23,265	1710	183	3	2140	140	14	2	5692	442	16	1
No. of Unique Transcripts	3456	117	5	0	10,322	504	0	0	6017	592	168	2	14,588	1551	175	3	1761	109	14	2	4273	376	16	1
No. of Unique Genes	1267	38	2	0	3509	161	0	0	2671	248	71	2	5965	645	83	2	1010	68	8	2	2295	201	9	1

[a]Pyr Only (pyrimidine-only)—triplex hits with only U or C in the Hoogsteen strand, U Only (uridine-only)—triplex hits with only U in the Hoogsteen strand, C Only (cytidine-only)—triplex hits with only C in the Hoogsteen strand.

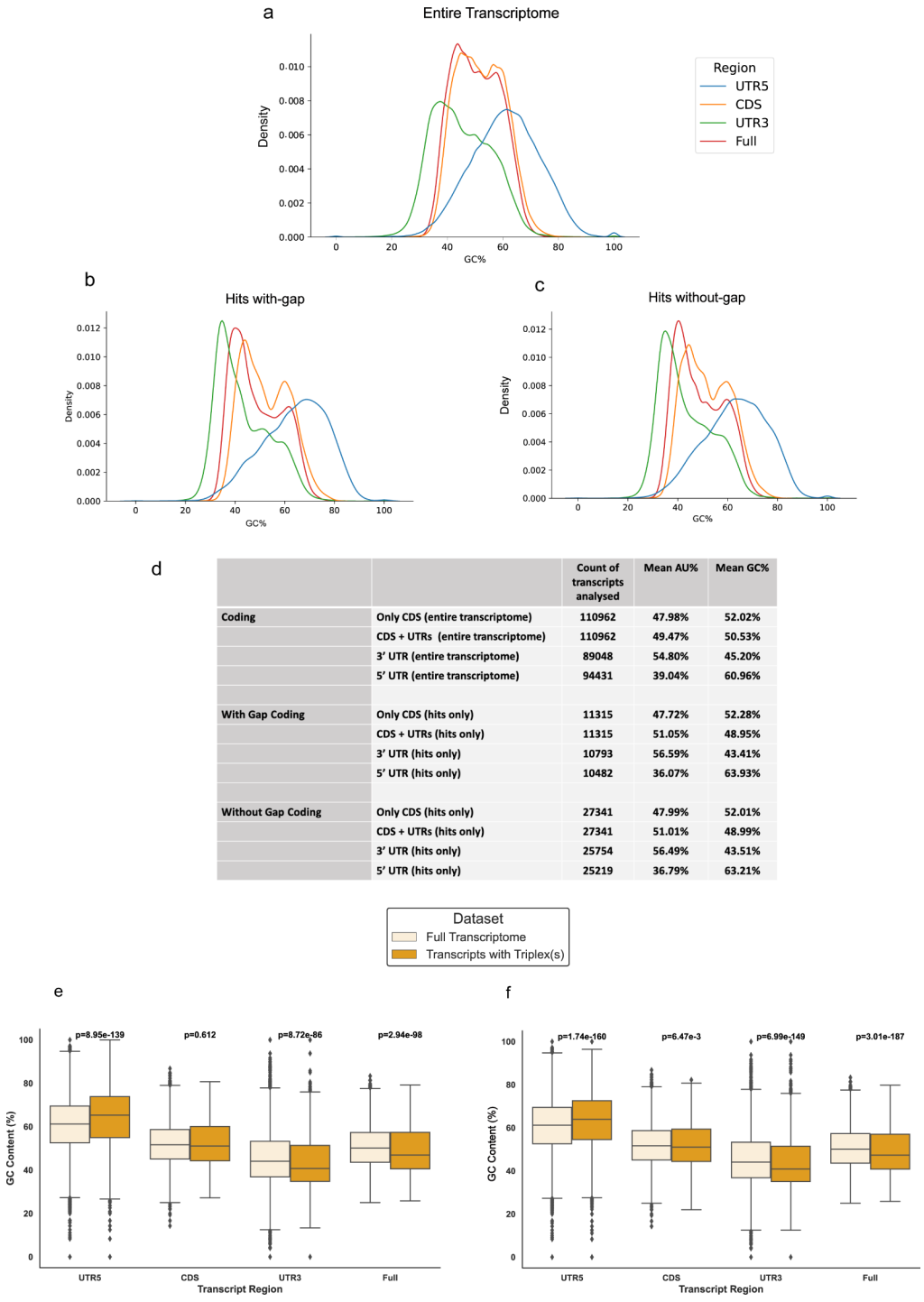
It is intuitive to expect these RNA structures to be present in the noncoding transcriptome, rather than in the coding transcriptome. Since lncRNAs are understood to be conserved on the basis of structure and not sequence, one may expect these structural elements to contribute toward the function of lncRNAs. However, the UTR regions of coding transcripts also fall under the category of regulatory elements and may be expected to contain RNA structural elements. For this reason, we analyzed the coding transcriptome of humans too, which yielded intriguing results. We observed that more than half the number of total hits of the coding region lay in the 3′ UTR.

Further, while in the noncoding portion of the transcriptome we observed that transcripts containing hits displayed a significantly higher AU usage, we additionally observed a significant difference in AU/GC usage in case of coding regions, too. In coding transcripts containing hits, GC usage was significantly higher in the 5′ UTR and significantly lower in the 3′ UTR than expected on average. (In this analysis, the triplexes that lie on the two boundary regions between the UTRs and the CDS have not been taken into account.) As for the CDS region, the small difference in GC usage is nonsignificant in the case of search in "with-gap" mode, and in the case of search "without-gap", the difference of 0.01% was significant with a p-value of 0.006. This indicates that while there seems to be a selection for transcripts with hits to have a preference toward a specific type of AU usage in the UTR regions, this preference is absent from the CDS region. This observation is in line with expectations,[66] and this GC usage bias in the UTRs points toward the occurrence of predicted intramolecular RNA triplex structures in the 5′ UTR and 3′ UTR of coding transcripts as being meaningful and biologically relevant. Just as RNA structures present in lncRNAs may be functional, the presence of such structures may contribute to the regulatory functions of the UTRs of coding transcripts. Since the CDS region is under evolutionary pressure to remain unaltered, such mutational changes may be less frequent and less observable.

Further, it is also intuitive that groups of genes, containing predicted triplexes and involved in particular common functions, may be found by analyzing the output of TRIPinRNA. The presence of these triplex structures may explain how some RNA−RNA or RNA−DNA interactions happen, and it may be interesting to explore cases like X-chromosome imprinting in further depth from the perspective of the presence of triplexes. Since the TRIPinRNA code runs (and gives output) chromosome-wise, the XCI triplexes were immediately apparent.

Some other predicted lncRNA hits of interest were MEG3, Airn, and KCNQ1OT1. Further, HELLPAR, MIR31HG, CYTOR, and PURPL lncRNAs showed a large number of predicted triplexes, which warrant further exploration. Within the coding subset of genes, hits with predicted triplexes in the 5′ UTR, for instance, are YY1, NFKB1, and FOXP1, in the CDS are DYNC1H1, NRG2, and MAP3K4 and in the 3′ UTR are DDX3Y and HNRNPA2B1. Examples of genes with predicted triplexes spanning the 5′ UTR and CDS region include MYCBP2 and LTBP3, and examples of ones spanning the CDS and 3′ UTR regions include SF1 and FBXL16. Lastly, it is observed that the 5′ UTR of the SOX2 transcript contains two predicted triplexes and that the SOX2-OT lncRNA contains an overwhelming number of predicted triplexes, too. The role of all of these and the numerous other triplexes could yield very interesting and valuable insights into the functions of these genes.

https://doi.org/10.1021/acs.biochem.4c00334
Biochemistry 2025, 64, 250−265



		Count of transcripts analysed	Mean AU%	Mean GC%
Coding	Only CDS (entire transcriptome)	110962	47.98%	52.02%
	CDS + UTRs (entire transcriptome)	110962	49.47%	50.53%
	3′ UTR (entire transcriptome)	89048	54.80%	45.20%
	5′ UTR (entire transcriptome)	94431	39.04%	60.96%
With Gap Coding	Only CDS (hits only)	11315	47.72%	52.28%
	CDS + UTRs (hits only)	11315	51.05%	48.95%
	3′ UTR (hits only)	10793	56.59%	43.41%
	5′ UTR (hits only)	10482	36.07%	63.93%
Without Gap Coding	Only CDS (hits only)	27341	47.99%	52.01%
	CDS + UTRs (hits only)	27341	51.01%	48.99%
	3′ UTR (hits only)	25754	56.49%	43.51%
	5′ UTR (hits only)	25219	36.79%	63.21%

Figure 8. Overview of GENCODE protein-coding data set results. Density plot of GC percent distribution in different portions of (a) the entire coding human transcriptome, (b) hit transcripts with predicted triplexes in "with-gap", and (c) hit transcripts with predicted triplexes in "without-gap" mode. (d) Tabulated number of transcripts analyzed along with their GC percentage. Difference in GC usage plotted with significance values noted (e) for run in "with-gap" mode and (f) for run in "without-gap" mode.

While these observations are exciting, they are not completely congruent with the prevailing understanding, which is that intramolecular triple helices must contain strands that are polypurine or poly pyrimidine-—only. Our observations in biophysical experiments have shown otherwise, however, and encouraged us to search computationally for this noncanonical type of triplexes.

Finally, it is observed that the triplexes predicted under the noncanonical category are also less thermally stable than the stability element-type triplexes. One possible explanation could be that these triplexes do not exist as constitutively formed structures but form and resolve with a temporality, based on external cues such as binding factors like proteins or small molecules. Further study into their formation and function will shed light on biological functions that are possibly equally unusual.

A comparison with other existing tools, Infernal and AlphaFold3 revealed that our script detected significantly

https://doi.org/10.1021/acs.biochem.4c00334
Biochemistry 2025, 64, 250−265

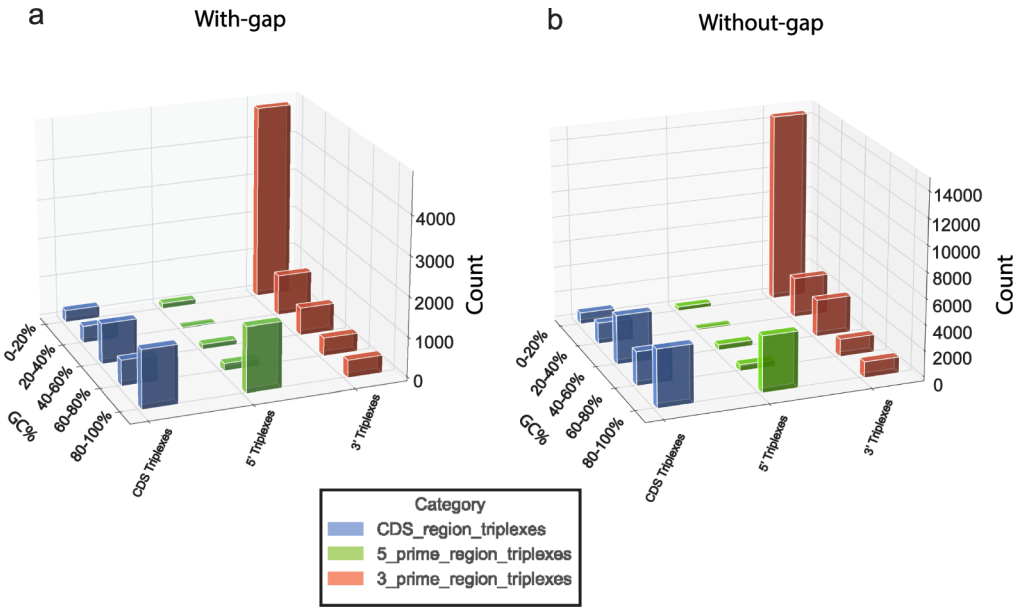


Figure 9. Distribution of the number of triplex hits with respect to region and GC percentage of the Hoogsteen strand of triplexes predicted from TRIPinRNA run in two modes: (a) with gap and (b) without gap.

more triplexes. Infernal uses covariance models to find RNA secondary structures by comparing input sequences to known RNA family models. It requires input sequences in the FASTA format along with multiple sequence alignments (MSAs) to build these models. Infernal was able to identify two known triple helices (MALAT1 and NEAT1) in the human genome using this model.[31] AlphaFold3, on the other hand, employs deep learning techniques to predict the folding of proteins and RNA based on sequence data. It accepts input sequences in FASTA format and outputs 3D structural predictions in PDB format. However, AlphaFold3 was unable to predict triple helix formation in RNA stretches (even smaller than 100 nt) that were obtained as hits from TRIPinRNA and hence was not suited for this specific purpose of finding potential triple helix-forming sequences ranging from around 100 to 1000 nt, in its current version.

In contrast, TRIPinRNA processed the entire transcriptome much more efficiently. Our script is well-suited for in-depth triplex structure prediction research, as it captures all the nuances of triplex patterns, providing a more focused analysis than other existing tools. It also provides detailed information and structured output in a CSV format, making it a powerful tool for triplex research. In the future, we envision being able to use the output obtained from TRIPinRNA to train machine learning software to detect triple helices with slightly varied features.

In conclusion, a study into the crystal structure of the noncanonical type of triplexes would be of abundant interest and value, but it is currently beyond the scope of our work. A successful analysis of such a crystal structure could shed light on the spatial arrangement of nucleotides, in effect opening up the understanding of the formation of such and other similar RNA structures.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.biochem.4c00334.

Sample output of TRIPinRNA (Figure S1); spatial distribution of 39 Xact triplexes (Figure S2); drawings of heteropyrimidine−purine strand-type triplexes (Figure S3); GC percent in "all" subset of noncoding hits (Figure S4); (a) GC percent in "coding hits", (b) GC percent "all" coding hits, CDS only, (c) GC percent "all" coding hits 5′ UTR, and (d) GC percent in "all" coding hits 3′ UTR only (Figure S5); statistical analyses for GC percent usage significance (Table S1); correlation plot for number of triplexes in a transcript with transcript length (Figure S6); (a) distribution of coding triplexes with GC usage (with-gap mode) and (b) distribution of coding triplexes with GC usage (without-gap mode) (Figure S7); thermal melting curves of triplex structures at 295 nm (Figure S8); crude density calculation of triple helix hits (Table S2) (PDF)

Triplex_all_with_gap_lncRNA (XLSX)

Triplex_all_no_gap_lncRNA (XLSX)

Triplex_all_with_gap_pcRNA (XLSX)

Triplex_all_no_gap_pcRNA (XLSX)

■ AUTHOR INFORMATION

Corresponding Author

Souvik Maiti − *CSIR-Institute of Genomics & Integrative Biology, Delhi 110025, India; Academy of Scientific and Innovative Research (AcSIR), Ghaziabad 201002, India; Institute of Genomics and Integrative Biology (IGIB)-National Chemical Laboratory (NCL) Joint Center, Council of Scientific and Industrial Research-NCL, Pune 411008, India;* orcid.org/0000-0001-9897-1419; Email: souvik@igib.res.in

Authors

Isha Rakheja − *CSIR-Institute of Genomics & Integrative Biology, Delhi 110025, India; Academy of Scientific and Innovative Research (AcSIR), Ghaziabad 201002, India*

Vishal Bharti − *CSIR-Institute of Genomics & Integrative Biology, Delhi 110025, India*

S Sahana − *CSIR-Institute of Genomics & Integrative Biology, Delhi 110025, India; Academy of Scientific and Innovative Research (AcSIR), Ghaziabad 201002, India*
Prosad Kumar Das − *CSIR-Institute of Genomics & Integrative Biology, Delhi 110025, India*
Gyan Ranjan − *CSIR-Institute of Genomics & Integrative Biology, Delhi 110025, India; Academy of Scientific and Innovative Research (AcSIR), Ghaziabad 201002, India*
Ajit Kumar − *CSIR-Institute of Genomics & Integrative Biology, Delhi 110025, India; Academy of Scientific and Innovative Research (AcSIR), Ghaziabad 201002, India*
Niyati Jain − *CSIR-Institute of Genomics & Integrative Biology, Delhi 110025, India;* orcid.org/0000-0002-1524-2411

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.biochem.4c00334

Author Contributions

[‖]I.R. and V.B. contributed equally to this work. S.M. and I.R. conceptualized the work; I.R. performed experiments with help from A.K. and N.J.; I.R. and V.B. wrote the manuscript; V.B., S.S., P.K.D., and G.R. performed the bioinformatics coding; and V.B. performed the bioinformatics-related analysis.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This study is funded by the Council for Scientific and Industrial Research (CSIR), Government of India, project MLP0139—"Targeting RNA Driven Processes: Novel Chemical Biology Approaches to Identify New Classes of RNA Modulators".

■ REFERENCES

(1) Derrien, T.; Johnson, R.; Bussotti, G.; Tanzer, A.; Djebali, S.; Tilgner, H.; Guernec, G.; Martin, D.; Merkel, A.; Knowles, D. G.; Lagarde, J.; Veeravalli, L.; Ruan, X.; Ruan, Y.; Lassmann, T.; Carninci, P.; Brown, J. B.; Lipovich, L.; Gonzalez, J. M.; Thomas, M.; Davis, C. A.; Shiekhattar, R.; Gingeras, T. R.; Hubbard, T. J.; Notredame, C.; Harrow, J.; Guigó, R. The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. *Genome Res.* **2012**, *22*, 1775−1789.
(2) Wang, C.; Wang, L.; Ding, Y.; Lu, X.; Zhang, G.; Yang, J.; Zheng, H.; Wang, H.; Jiang, Y.; Xu, L. LncRNA Structural Characteristics in Epigenetic Regulation. *Int. J. Mol. Sci.* **2017**, *18*, 2659.
(3) Dong, P.; Xiong, Y.; Yue, J.; Hanley, S. J. B.; Kobayashi, N.; Todo, Y.; Watari, H. Long non-coding RNA NEAT1: A novel target for diagnosis and therapy in human tumors. *Front. Genet.* **2018**, *9*, 471.
(4) Arora, A.; Maiti, S. Effect of loop orientation on quadruplex - TMPyP4 interaction. *J. Phys. Chem. B* **2008**, *112*, 8151−8159.
(5) Cash, D. D.; Cohen-Zontag, O.; Kim, N. K.; Shefer, K.; Brown, Y.; Ulyanov, N. B.; Tzfati, Y.; Feigon, J. Pyrimidine motif triple helix in the Kluyveromyces lactis telomerase RNA pseudoknot is essential for function in vivo. *Proc. Natl. Acad. Sci. U. S. A.* **2013**, *110*, 10970−10975.
(6) Brown, J. A.; Kinzig, C. G.; Degregorio, S. J.; Steitz, J. A. Methyltransferase-like protein 16 binds the 3′-terminal triple helix of MALAT1 long noncoding RNA. *Proc. Natl. Acad. Sci. U. S. A.* **2016**, *113*, 14013−14018.
(7) Wilusz, J. E.; Freier, S. M.; Spector, D. L. 3′End Processing of a Long Nuclear-Retained Noncoding RNA Yields a tRNA-like Cytoplasmic RNA. *Cell* **2008**, *135*, 919−932.
(8) Xing, Y. H.; Chen, L. L. Processing and roles of snoRNA-ended long noncoding RNAs. In *Critical Reviews in Biochemistry and Molecular Biology*; Taylor and Francis Ltd., 2018.
(9) Szabat, M.; Kierzek, E.; Kierzek, R. Modified RNA triplexes: Thermodynamics, structure and biological potential. *Sci. Rep.* **2018**, *8* (1), 13023.
(10) Agarwal, T.; Jayaraj, G.; Pandey, S. P.; Agarwala, P.; Maiti, S. RNA G-Quadruplexes: G-quadruplexes with "U"Turns. *Curr Pharm Des.* **2012**, *18*, 2102−2111.
(11) Ghosh, A.; Pandey, S. P.; Joshi, D. C.; Rana, P.; Ansari, A. H.; Sundar, J. S.; Singh, P.; Khan, Y.; Ekka, M. K.; Chakraborty, D.; Maiti, S. Identification of G-quadruplex structures in MALAT1 lncRNA that interact with nucleolin and nucleophosmin. *Nucleic Acids Res.* **2023**, *51*, 9415−9431.
(12) Felsenfeld, G.; Davies, D. R.; Rich, A. FORMATION OF A THREE STRANDED POLYNUCLEOTIDE MOLECULE. *J. Am. Chem. Soc.* **1957**, *79*, 2023−2024.
(13) Brown, J. A.; Valenstein, M. L.; Yario, T. A.; Tycowski, K. T.; Steitz, J. A. Formation of triple-helical structures by the 3′-end sequences of MALAT1 and MEN*β* noncoding RNAs. *Proc. Natl. Acad. Sci. U. S. A.* **2012**, *109*, 19202−19207.
(14) Wilusz, J. E.; JnBaptiste, C. K.; Lu, L. Y.; Kuhn, C. D.; Joshua-Tor, L.; Sharp, P. A. A triple helix stabilizes the 3′ends of long noncoding RNAs that lack poly(A) tails. *Genes Dev.* **2012**, *26*, 2392−2407.
(15) Ageeli, A. A.; McGovern-Gooch, K. R.; Kaminska, M. M.; Baird, N. J. Finely tuned conformational dynamics regulate the protective function of the lncRNA MALAT1 triple helix. *Nucleic Acids Res.* **2019**, *47*, 1468−1481.
(16) Brown, J. A.; Kinzig, C. G.; Degregorio, S. J.; Steitz, J. A. Hoogsteen-position pyrimidines promote the stability and function of the MALAT1 RNA triple helix. *RNA* **2016**, *22*, 743−749.
(17) Brown, J. A.; Bulkley, D.; Wang, J.; Valenstein, M. L.; Yario, T. A.; Steitz, T. A.; Steitz, J. A. Structural insights into the stabilization of MALAT1 noncoding RNA by a bipartite triple helix. *Nat. Struct. Mol. Biol.* **2014**, *21*, 633−640.
(18) Theimer, C. A.; Blois, C. A.; Feigon, J. Structure of the human telomerase RNA pseudoknot reveals conserved tertiary interactions essential for function. *Mol. Cell* **2005**, *17*, 671−682.
(19) Ulyanov, N. B.; Shefer, K.; James, T. L.; Tzfati, Y. Pseudoknot structures with conserved base triples in telomerase RNAs of ciliates. *Nucleic Acids Res.* **2007**, *35*, 6150−6160.
(20) Conrad, N. K. The emerging role of triple helices in RNA biology. *Wiley Interdisciplinary Rev.: RNA* **2014**, *5* (1), 15−29.
(21) Donlic, A.; Zafferani, M.; Padroni, G.; Puri, M.; Hargrove, A. E. Regulation of MALAT1 triple helix stability and in vitro degradation by diphenylfurans. *Nucleic Acids Res.* **2020**, *48*, 7653−7664.
(22) Donlic, A.; Morgan, B. S.; Xu, J. L.; Liu, A.; Roble, C.; Hargrove, A. E. Discovery of Small Molecule Ligands for MALAT1 by Tuning an RNA-Binding Scaffold. *Angew. Chem., Int. Ed.* **2018**, *57*, 13242−13247.
(23) Warner, K. D.; Hajdin, C. E.; Weeks, K. M. Principles for targeting RNA with drug-like small molecules. *Nat. Rev. Drug Discovery* **2018**, *17*, 547−558.
(24) Moser, H. E.; Dervan, P. B. Sequence-specific cleavage of double helical DNA by triple helix formation. *Science* **1987**, *238*, 645−650.
(25) Abulwerdi, F. A.; Xu, W.; Ageeli, A. A.; Yonkunas, M. J.; Arun, G.; Nam, H.; Schneekloth, J. S., Jr; Dayie, T. K.; Spector, D.; Baird, N.; et al. Selective Small-Molecule Targeting of a Triple Helix Encoded by the Long Noncoding RNA, MALAT1. *ACS Chem. Biol.* **2019**, *14* (2), 223−235.
(26) Falese, J. P.; Donlic, A.; Hargrove, A. E. Targeting RNA with small molecules: From fundamental principles towards the clinic. In *Chemical Society Reviews*; Royal Society of Chemistry, 2021.
(27) Nawrocki, E. P.; Eddy, S. R. Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics* **2013**, *29*, 2933−2935.
(28) Tycowski, K. T.; Shu, M. D.; Steitz, J. A. Myriad Triple-Helix-Forming Structures in the Transposable Element RNAs of Plants and Fungi. *Cell Rep.* **2016**, *15*, 1266−1276.
(29) Torabi, S. F.; Vaidya, A. T.; Tycowski, K. T.; DeGregorio, S. J.; Wang, J.; Shu, M. D.; Steitz, T. A.; Steitz, J. A. RNA stabilization by a poly(A) tail 3′-end binding pocket and other modes of poly(A)-RNA interaction. *Science* **2021**, *371* (6529), eabe6523.
(30) Torabi, S.-F.; Chen, Y.-L.; Zhang, K.; Wang, J.; DeGregorio, S. J.; Vaidya, A. T.; Su, Z.; Pabit, S. A.; Chiu, W.; Pollack, L.; et al. Structural

analyses of an RNA stability element interacting with poly(A). *Proc. Natl. Acad. Sci. U. S. A.* **2021**, *118* (14), No. e2026656118.
(31) Zhang, B.; Mao, Y. S.; Diermeier, S. D.; Novikova, I. V.; Nawrocki, E. P.; Jones, T. A.; Lazar, Z.; Tung, C. S.; Luo, W.; Eddy, S. R.; Sanbonmatsu, K. Y.; Spector, D. L. Identification and Characterization of a Class of MALAT1-like Genomic Loci. *Cell Rep.* **2017**, *19*, 1723−1738.
(32) Lexa, M.; Martínek, T.; Burgetová, I.; Kopeček, D.; Brázdová, M. A dynamic programming algorithm for identification of triplex-forming sequences. *Bioinformatics* **2011**, *27*, 2510−2517.
(33) Hon, J.; Martínek, T.; Rajdl, K.; Lexa, M. Triplex: An R/Bioconductor package for identification and visualization of potential intramolecular triplex patterns in DNA sequences. *Bioinformatics* **2013**, *29*, 1900−1901.
(34) Holder, I. T.; Wagner, S.; Xiong, P.; Sinn, M.; Frickey, T.; Meyer, A.; Hartig, J. S. Intrastrand triplex DNA repeats in bacteria: a source of genomic instability. *Nucleic Acids Res.* **2015**, *43*, 10126−10142.
(35) Hoyne, P. R.; Edwards, L. M.; Viari, A.; Maher, L. J. 3rd. Searching genomes for sequences with the potential to form intrastrand triple helices. *J. Mol. Biol.* **2000**, *302*, 797−809.
(36) Cer, R. Z.; Bruce, K. H.; Mudunuri, U. S.; Yi, M.; Volfovsky, N.; Luke, B. T.; Bacolla, A.; Collins, J. R.; Stephens, R. M. Non-B DB: a database of predicted non-B DNA-forming motifs in mammalian genomes. *Nucleic Acids Res.* **2011**, *39*, D383−91.
(37) Pandey, S.; Agarwala, P.; Maiti, S. Effect of loops and G-quartets on the stability of RNA G-quadruplexes. *J. Phys. Chem. B* **2013**, *117*, 6896−6905.
(38) Lee, H.-T.; Carr, C. E.; Khutsishvili, I.; Marky, L. A. Effect of Loop Length and Sequence on the Stability of DNA Pyrimidine Triplexes with TAT Base Triplets. *J. Phys. Chem. B* **2017**, *121*, 9175−9184.
(39) Berselli, M.; Lavezzo, E.; Toppo, S. NeSSie: a tool for the identification of approximate DNA sequence symmetries. *Bioinformatics* **2018**, *34*, 2503−2505.
(40) Zhao, Y.; Li, H.; Fang, S.; Kang, Y.; Wu, W.; Hao, Y.; Li, Z.; Bu, D.; Sun, N.; Zhang, M. Q.; Chen, R. NONCODE 2016: An informative and valuable data source of long non-coding RNAs. *Nucleic Acids Res.* **2016**, *44*, D203−D208.
(41) Rakheja, I.; Ansari, A. H.; Ray, A.; Chandra Joshi, D.; Maiti, S. Small molecule quercetin binds MALAT1 triplex and modulates its cellular function. *Mol. Ther. Nucleic Acids* **2022**, *30*, 241−256.
(42) Ruszkowska, A.; Ruszkowski, M.; Hulewicz, J. P.; Dauter, Z.; Brown, J. A. Molecular structure of a U●A-U-rich RNA triple helix with 11 consecutive base triples. *Nucleic Acids Res.* **2020**, *48*, 3304−3314.
(43) Frankish, A.; Diekhans, M.; Jungreis, I.; Lagarde, J.; Loveland, J. E.; Mudge, J. M.; Sisu, C.; Wright, J. C.; Armstrong, J.; Barnes, I.; et al. GENCODE 2021. *Nucleic Acids Res.* **2021**, *49* (D1), D916−D923.
(44) Kuang, Y.; Shen, W.; Zhu, H.; Huang, H.; Zhou, Q.; Yin, W.; Zhou, Y.; Cao, Y.; Wang, L.; Li, X.; Ren, C.; Jiang, X. The role of lncRNA just proximal to XIST (JPX) in human disease phenotypes and RNA methylation: The novel biomarker and therapeutic target potential. *Biomed. Pharmacother.* **2022**, *155*, 113753.
(45) Swain, M.; Ageeli, A. A.; Kasprzak, W. K.; Li, M.; Miller, J. T.; Sztuba-Solinska, J.; Schneekloth, J. S.; Koirala, D.; Piccirili, J.; Fraboni, A. J.; Murelli, R. P.; Wlodawer, A.; Shapiro, B. A.; Baird, N.; Le Grice, S. F. J. Dynamic bulge nucleotides in the KSHV PAN ENE triple helix provide a unique binding platform for small molecule ligands. *Nucleic Acids Res.* **2021**, *49*, 13179−13193.
(46) Tiwari, R.; Haque, L.; Bhuiya, S.; Das, S. Third strand stabilization of poly(U)·poly(A)* poly(U) triplex by the naturally occurring flavone luteolin: A multi-spectroscopic approach. *Int. J. Biol. Macromol.* **2017**, *103*, 692−700.
(47) Motosugi, N.; Okada, C.; Sugiyama, A.; Kawasaki, T.; Kimura, M.; Shiina, T.; Umezawa, A.; Akutsu, H.; Fukuda, A. Deletion of lncRNA XACT does not change expression dosage of X-linked genes, but affects differentiation potential in hPSCs. *Cell Rep.* **2021**, *35*, 109222.
(48) Yang, F.; Deng, X.; Ma, W.; Berletch, J. B.; Rabaia, N.; Wei, G.; Moore, J. M.; Filippova, G. N.; Xu, J.; Liu, Y.; et al. The lncRNA Firre anchors the inactive X chromosome to the nucleolus by binding CTCF and maintains H3K27me3 methylation. *Genome Biol.* **2015**, *52*, 16.
(49) Shi, X.; Cui, Z.; Liu, X.; Wu, S.; Wu, Y.; Fang, F.; Zhao, H. LncRNA FIRRE is activated by MYC and promotes the development of diffuse large B-cell lymphoma via Wnt/β-catenin signaling pathway. *Biochem. Biophys. Res. Commun.* **2019**, *510*, 594−600.
(50) Engreitz, J. M.; Pandya-Jones, A.; McDonel, P.; Shishkin, A.; Sirokman, K.; Surka, C.; Kadri, S.; Xing, J.; Goren, A.; Lander, E. S.; et al. The Xist lncRNA exploits three-dimensional genome architecture to spread across the X chromosome. *Science* **2013**, *341* (6147), 1237973.
(51) Vigneau, S.; Augui, S.; Navarro, P.; Avner, P.; Clerc, P. An essential role for the DXPas34 tandem repeat and Tsix transcription in the counting process of X chromosome inactivation. *Proc. Natl. Acad. Sci. U. S. A.* **2006**, *103*, 7390−7395.
(52) Hosoi, Y.; Soma, M.; Shiura, H.; Sado, T.; Hasuwa, H.; Abe, K.; Kohda, T.; Ishino, F.; Kobayashi, S. Female mice lacking Ftx lncRNA exhibit impaired X-chromosome inactivation and a microphthalmia-like phenotype. *Nat. Commun.* **2018**, *9* (1), 3829.
(53) Zuker, M. Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res.* **2003**, *31*, 3406−3415.
(54) Mergny, J. L.; Li, J.; Lacroix, L.; Amrane, S.; Chaires, J. B. Thermal difference spectra: A specific signature for nucleic acid structures. *Nucleic Acids Res.* **2005**, *33* (16), No. e138−e138.
(55) Annala, M.; Taavitsainen, S.; Vandekerkhove, G.; Bacon, J. V. W.; Beja, K.; Chi, K. N.; Nykter, M.; Wyatt, A. W. Frequent mutation of the FOXA1 untranslated region in prostate cancer. *Commun. Biol.* **2018**, *1* (1), 122.
(56) Lim, Y.; Arora, S.; Schuster, S. L.; Corey, L.; Fitzgibbon, M.; Wladyka, C. L.; Wu, X.; Coleman, I. M.; Delrow, J. J.; Corey, E.; True, L. D.; Nelson, P. S.; Ha, G.; Hsieh, A. C. Multiplexed functional genomic analysis of 5' untranslated region mutations across the spectrum of prostate cancer. *Nat. Commun.* **2021**, *12* (1), 4217.
(57) Leppek, K.; Das, R.; Barna, M. Functional 5′ UTR mRNA structures in eukaryotic translation regulation and how to find them. *Nat. Rev. Mol. Cell Biol.* **2018**, *19*, 158−174.
(58) Mayya, V. K.; Duchaine, T. F. Ciphers and executioners: How 3′-untranslated regions determine the fate of messenger RNAs. *Front. Genet.* **2019**, *10*, 6.
(59) Kunkler, C. N.; Hulewicz, J. P.; Hickman, S. C.; Wang, M. C.; McCown, P. J.; Brown, J. A. Stability of an RNA●DNA-DNA triple helix depends on base triplet composition and length of the RNA third strand. *Nucleic Acids Res.* **2019**, *47*, 7213−7222.
(60) Brown, J. A. Unraveling the structure and biological functions of RNA triple helices. *Wiley Interdiscip. Rev.: RNA.* **2020**, *11*, No. e1598.
(61) Howard, F. B.; Miles, H. T.; Ross, P. D. The poly-(dT).Cntdot.2Poly(dA) triple helix. *Biochemistry* **1995**, *34*, 7135−7144.
(62) Gondeau, C.; Maurizot, J. C.; Durand, M. Circular dichroism and UV melting studies on formation of an intramolecular triplex containing parallel T*A: T and G*G: C triplets: netropsin complexation with the triplex. *Nucleic Acids Res.* **1998**, *26*, 4996−5003.
(63) Gondeaut, C.; Maurizot, J. C.; Durand, M. Spectroscopic studies on ethidium bromide binding to intramolecular parallel and antiparallel triple helices containing T*A: T and G*G: C triplets. *J. Biomol. Struct. Dyn.* **2000**, *17*, 879−886.
(64) Lee, J. S.; Woodsworth, M. L.; Latimer, L. J.; Morgan, A. R. Poly(pyrimidine). poly(purine) synthetic DNAs containing 5-methylcytosine form stable triplexes at neutral pH. *Nucleic Acids Res.* **1984**, *12*, 6603−6614.
(65) He, Y.; Scaria, P. V.; Shafer, R. H. Studies on formation and stability of the d[G(AG)5]* d[G(AG)5]. d[C(TC)5] and d[G(TG)-5]* d[G(AG)5]. d[C(TC)5] triple helices. *Biopolymers* **1997**, *41*, 431−441.
(66) Louie, E.; Ott, J.; Majewski, J. Nucleotide frequency variation across human genes. *Genome Res.* **2003**, *13*, 2594−2601.

https://doi.org/10.1021/acs.biochem.4c00334
Biochemistry 2025, 64, 250−265