Genome-Wide Dual-Selection Unveils Novel Self-Cleaving Ribozymes in the Human Genome

Zhe Zhang[1], Jian Zhan[1,2,3*] and Yaoqi Zhou[1*]

[1]Institute for Systems and Physical Biology, Shenzhen Bay Laboratory, Shenzhen 518107, China
[2]Ribopeutic (Shenzhen) Co., Ltd., Futian, Shenzhen, Guangdong Province, 518000, China
[3]Ribopeutic Inc., Qiantang, Hangzhou, Zhejiang Province, 310018, China
[*] Co-corresponding authors (zhanjian@szbl.ac.cn and zhouyq@szbl.ac.cn)

Abstract

The landscape of catalytic RNA in complex eukaryotes remains poorly charted. Although self-cleaving ribozymes are widespread in microbial and viral genomes, their existence and functional roles in humans remain limited. Here, we introduce a generalizable, genome-wide discovery platform that integrates high-throughput signals from two complementary adapter ligation assays—3P-seq and 5OH-seq—to specifically capture RNA fragments bearing cleavage signatures (2′,3′-cyclic phosphate/3′-phosphate and 5′-hydroxyl termini). By applying a dedicated computational scoring algorithm to human genomic data, we systematically identified four previously unrecognized self-cleaving ribozymes. These ribozymes localize to diverse genomic features: an exon of WDFY1, an intron of PLD5 embedded within a repetitive element, an LTR retrotransposon, and the antisense strand of an SYNJ2BP intron. Truncation analyses defined minimal functional cores of ~70–100 nucleotides, and computational modeling suggests that they adopt novel structural architectures. Our work establishes a powerful strategy for transcriptome-wide mining of catalytic RNA, expands the catalogue of human ribozymes from 6 to 10, and reveals new layers of functional complexity in the human RNA world.

Main

Self-cleaving Ribozymes are an important class of non-coding RNAs with catalytic functions[1] that catalyze the site-specific self-cleavage of phosphodiester backbone[2]. Recent advancements in computational and high-throughput techniques have significantly accelerated the pace of ribozyme discovery[3,4]. However, the vast majority of self-cleaving ribozymes identified to date originate from lower organisms. Only 6 self-cleavage ribozymes are found in human so far (hammerhead ribozymes (HH9, HH10) located with introns[5], the HDV-like cytoplasmic polyadenylation element-binding protein 3 (CPEB3) ribozyme[6], 2 twister sister-like (TS-like) ribozymes[7] (LINE-1 and OR4K15), and the hovlinc ribozyme found in very long intergenic non-coding RNAs (vlincRNAs)[8]. Although these self-cleaving ribozymes lack structural similarity, they all accomplish site-specific

cleavage through a nucleophilic attack by a 2′-oxygen on the adjacent phosphodiester bond. This reaction produces cleavage termini characterized by a 2′,3′-cyclic phosphate group and a 5′-hydroxyl group, respectively.

Previously, the Kapranov group utilized the RppH 5′ pyrophosphohydrolase and the XRN-1 5′-3′ exoribonuclease to degrade the other transcriptional products, except for the RNA fragments containing the 5′-hydroxyl. This led to the discovery of the hovlinc ribozyme among 28 candidates. Although the low discovery rate of self-cleaving ribozymes suggests the rarity in complex organisms, the question of whether other self-cleaving ribozymes exist in the human genome merits further study.

To probe the genome for potential self-cleaving ribozymes, we developed a strategy that leverages the unique specificity of RtcB ligase, an enzyme that joins a 5′-hydroxyl terminus to a 3′-phosphate or 2′,3′-cyclic phosphate terminus [9]. This was achieved by utilizing RtcB ligase in conjunction with specifically designed adapter RNAs to capture RNA molecules bearing these characteristic ends generated from self-cleavage. Among 96 candidate sequences subjected to experimental verification, six were confirmed to exhibit self-cleaving activity. Two of these active sequences had been documented previously, whereas the other four represent novel self-cleaving ribozymes. It is noteworthy that one self-cleaving ribozyme was identified within an exonic region of the WDFY1 (WD repeat and FYVE domain containing 1) protein, suggesting that it may play a direct role in regulating protein expression. To the best of our knowledge, this is the first report of a self-cleaving ribozyme residing in a protein-coding exon, and thus, significantly expands our understanding of the functional repertoire and mechanistic diversity of self-cleaving ribozymes.

Results

Genome-wide discovery of self-cleaving ribozymes

Genomic libraries were prepared using the Illumina tagmentation kit, which employs a Tn5 transposase to simultaneously fragment and tagmate DNA, streamlining the library construction process. Briefly, the tagmentation reaction fragmented genomic DNA and ligated Nextera s7 and Nextera s5 adapter sequences to both ends of the fragments in a single step. Then, the termini of the genomic library fragments were appended with T7 promoter through PCR amplification, facilitating

downstream in vitro transcription (Fig. 1a). These RNA transcripts along with self-cleavage products obtained from the genomic library offers the distinct advantage of being unaffected by intracellular proteases. Unlike prior methods[8] that solely enriched for 5′-hydroxyl ends from self-cleavage, our approach specifically captures cleavage products containing either 2′,3′-cyclic phosphate or 5′-hydroxyl termini from self-cleavage. The 3P-seq is designed to isolate RNA fragments possessing 2′,3′-cyclic phosphate or 3′-phosphate ends. These 3′-ends were joined with a synthetic RNA linker that features a 5′-hydroxyl group by taking advantage of the unique ligation activity of RtcB ligase. Meanwhile, the 5OH-seq captures RNAs with 5′-hydroxyl termini by employing RtcB ligates these 5′-hydroxyl ends to a separate RNA linker that itself carries a 3′-phosphate group. A desthiobiotin modification was strategically engineered at the distal ends of both RNA linkers with a dual purpose: first, to minimize the formation of off-target ligation products, and second, to enable highly specific capture and enrichment by using streptavidin-based purification methods. The enriched products were subsequently processed into sequencing libraries for high-throughput analysis.

To minimize false positives, we employed paired-end sequencing and implemented a stringent quality filter based on the Mapping Quality (MAPQ) score from the SAM files, retaining only alignments with a MAPQ $\geqslant$ 30, which is widely regarded as a threshold for high-confidence mappings. The precise cleavage sites were inferred from the alignment boundaries, corresponding to the junctions between the inserted fragments and the adapter sequences. The distribution of these putative cleavage sites was then quantified by counting the supporting reads. During data processing, we observed that a substantial number of cleavage sites remained clustered within specific genomic regions. Given that self-cleaving ribozymes typically catalyze hydrolysis at unique sites, we rationalized that sites supported by an exceptionally low number of reads within these clusters were likely artifacts, potentially resulting from random RNA degradation. Consequently, a sophisticated denoising step was employed, which eliminated sites based on an intra-window signal intensity ratio. Furthermore, to enhance the reliability of our findings, we integrated the results from two independent experiments—3P-seq and 5OH-seq. A composite scoring system was designed to prioritize cleavage sites detected in both assays, under the premise that such "consensus" sites exhibit a lower probability of being false positives.

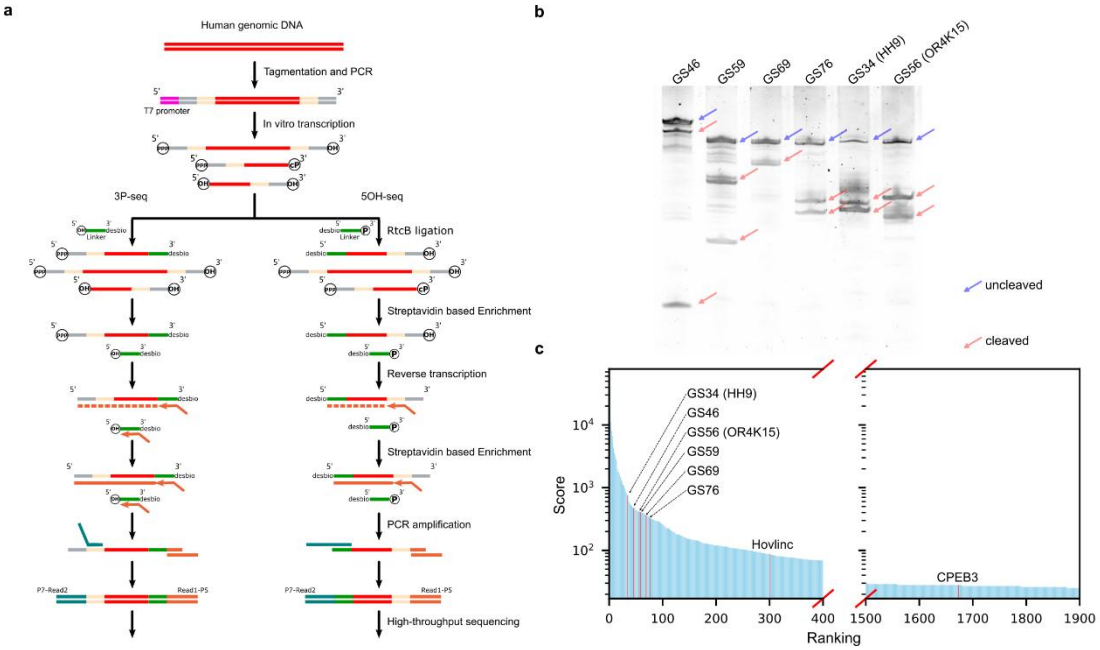

Fig. 1: Genome-scale discovery of self-cleaving ribozymes in human genome. a, The discovery flowchart: To profile RNA cleavage products, we generated a protease-resistant genomic RNA library through in vitro transcription. RtcB ligase and terminally modified RNA linkers were utilized to selectively capture and distinguish fragments bearing 2′,3′-cyclic phosphate or 5′-OH ends, followed by streptavidin bead enrichment and sequencing library construction. b, Urea PAGE analysis of the six experimentally validated ribozyme candidates from the top 96 hits. Hit nomenclature reflects its ranking (e.g., GS46 corresponds to the hit ranked 46th). c, Scoring distribution of the top 1,900 hits. Hits corresponding to known ribozymes and novel ribozymes that have been validated through experiments are explicitly marked in red.

Self-cleaving ribozymes identified in HeLa genome

Applying this integrated approach, we successfully identified four out of the five known self-cleaving ribozymes, with HH9 and OR4K15 ranking within the top 96 hits (Fig. 1b). However, the hovlinc ribozyme, discovered more recently, was ranked 301st (Fig. 1c), primarily due to its low detection count (2 reads) in the 5OH-seq assay. The CPEB3 ribozyme was ranked 1,673rd (Fig. 1c) and was not detected in the 5′OH-seq assay.

To gain some understanding of low priority in hovlinc and CPEB3 ribozymes, we generated normalized coverage files from the BAM files and visualized the signal intensity across the targeted gene regions by using the deepTools suite[10]. Analysis of the signal profiles across identified

ribozymes shows a spectrum: from those with a pronounced signal asymmetry (Fig. 2a, 2c, 2f, 2g and 2h) to those with relatively balanced signals at both ends (Fig. 2b, 2d and 2e). This variation is principally dictated by the unique sequence and the consequent secondary/tertiary structure of each ribozyme, which directly affects the mapping efficacy of the RNA linkers during the library preparation. These inherent structural differences are a major source of the significant variation in observed sequencing depth. For hovlinc and CPEB3 ribozymes, their characteristically stable 3′ terminal structures likely impede efficient adapter ligation, resulting in negligible 3′ end capture signal and thus a substantially lower ranking in our analysis, highlighting a specific limitation in detecting self-cleaving ribozymes with such structural features.

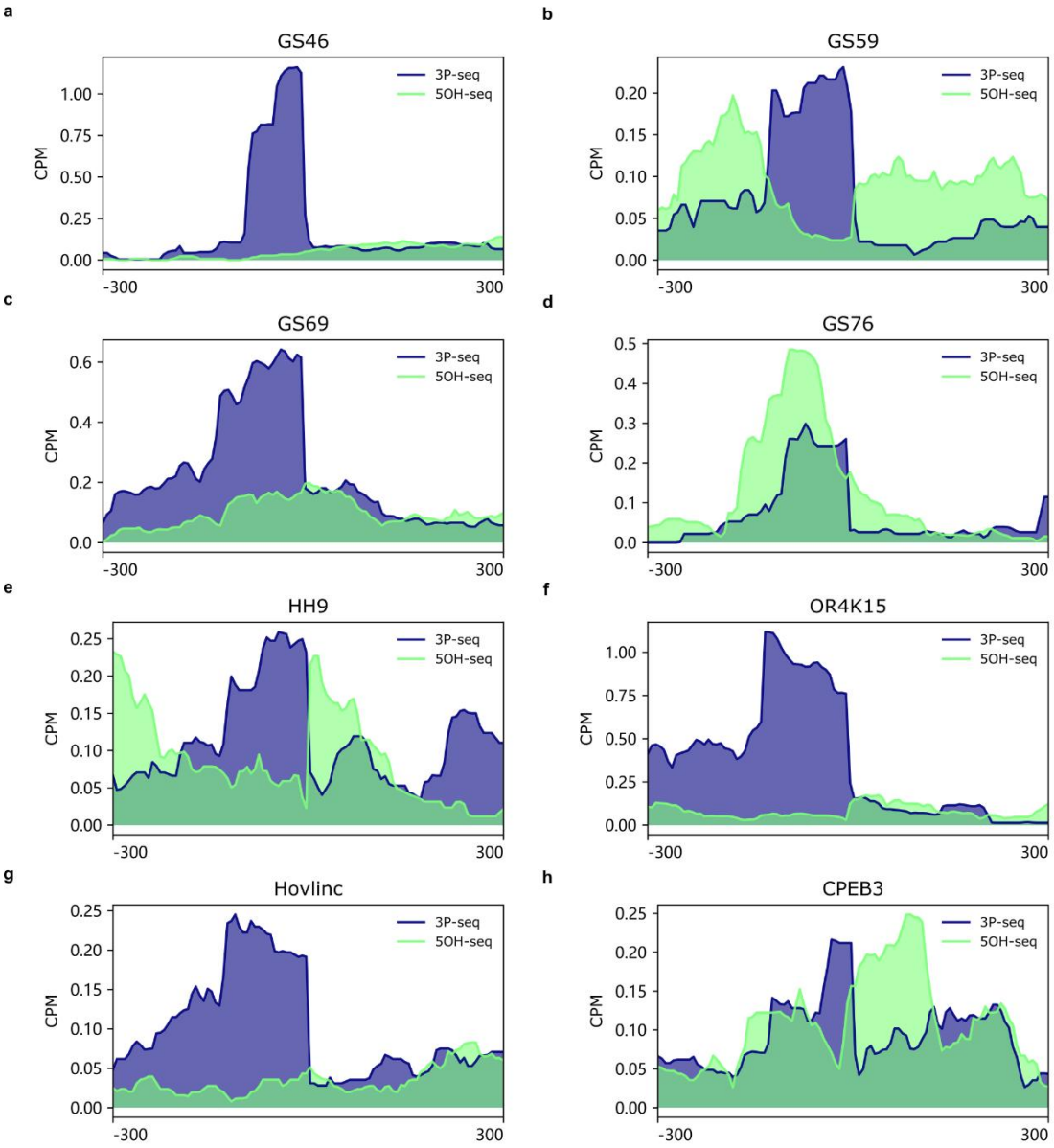

Fig. 2: Normalized signal from the two genome-wide selection assays. The y-axis represents the counts per million (CPM) of mapped reads, while the x-axis corresponds to a 600-nucleotide (nt) genomic window centered on the cleavage site. a–d illustrate four novel ribozymes identified in this study. e and f depict two previously reported ribozymes (HH9 ranked 34th, OR4K15 ranked 56th) that were confirmed among the top 96 validated sequences in our screening. In contrast, g and h show two known ribozymes that ranked lower in our results, indicating their relatively lower abundance or activity under the experimental conditions.

New ribozymes with potential novel function

We characterized four novel self-cleaving ribozymes (GS46, GS59, GS69, GS76), each occupying a distinct genomic locus that suggests unique regulatory potentials. Particularly notable is GS46, which was located within the eighth exon of the protein-coding gene WDFY1 (WD repeat and FYVE domain containing 1); its self-cleavage activity is thus supposed to truncate the host mRNA, potentially triggering nonsense-mediated decay and leading to reduced expression of the WDFY1 protein. Furthermore, GS59 and GS69 were identified within repetitive genomic architectures: GS59 was situated in the first intron of PLD5 (Phospholipase D Family Member 5) and embedded within a DNA repeat element, while GS69 was found within an intron of a novel transcript that itself resides in a Long Terminal Repeat (LTR) retrotransposon. Additionally, GS76 was mapped to the reverse (antisense) strand of the first intron of the SYNJ2BP (Synaptojanin 2 Binding Protein) gene, an intriguing location that raises the possibility of its involvement in regulating the sense transcript.

To elucidate the structural basis of the self-cleaving activity, we conducted a preliminary investigation of the four newly identified ribozymes. It is important to note that the RNA fragments captured in our sequencing libraries do not necessarily represent the minimal structural module required for catalytic activity as we demonstrated previously[7]. Therefore, to delineate the core functional domain of each ribozyme, we systematically generated a series of truncated variants. Given that the 3P-seq assay yielded stronger signals for most of these ribozymes (Fig. 2), indicating a more defined or stable 3′ end post-cleavage, the truncation strategy primarily involved progressive shortening from the 3′ terminus. The catalytic competence of these truncated constructs was then assessed by analyzing their self-cleavage efficiency via polyacrylamide gel

electrophoresis (PAGE). This empirical approach allowed us to identify the minimal functional core for each ribozyme, with optimized lengths of approximately 70 nt, 100 nt, 80 nt, and 80 nt, respectively (Fig. 3a). Subsequent computational prediction of the secondary and tertiary structures for these minimized variants suggested that each adopts a distinct and previously uncharacterized structural fold (Fig. 3b), potentially representing novel architectural classes among self-cleaving ribozymes.

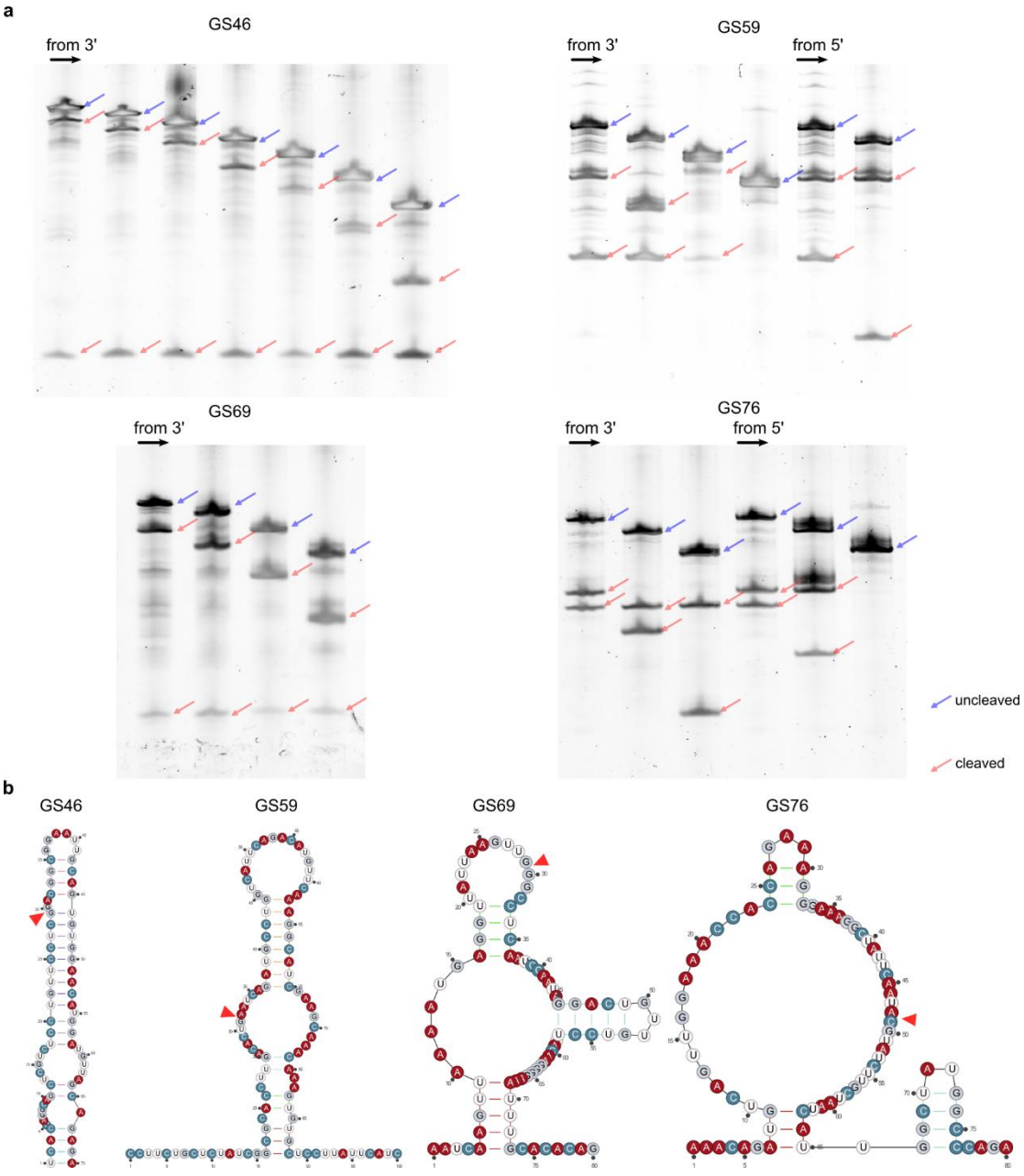

Fig. 3: Truncations and predicted secondary structures of the 4 new ribozymes. a, PAGE analysis of the truncated variants of the 4 new ribozymes. A series of truncated variants was generated from the original sequence for activity assay. The 3′ end (or 5′ end) was anchored, and sequential 20-nt

deletions were made from the opposite terminus. This process continued until the whole fragment reached a minimum length of 100 nt, which was determined to be optimal for efficient molecular cloning. b, Predicted secondary structure of the minimal truncated ribozymes. The secondary structures of the RNA sequences were predicted by employing the computational tool SPOT-RNA[11]. The red triangle indicates the cleavage site.

Discussion

Self-cleaving ribozymes are ubiquitous in lower organisms, yet their representation in complex eukaryotic genomes remains enigmatic. Although prior studies have identified a handful of examples in humans—including the HDV-like CPEB3 ribozyme[6], two hammerhead ribozymes[5], and the hovlinc ribozyme[8]—these discoveries relied heavily on known sequence motifs， specific biochemical properties or genome-scale search according to the discovery of hovlinc ribozyme[8]. This raises a fundamental question: is the scarcity of known human ribozymes biological reality, or a reflection of methodological constraints?

To address this, we developed a high-throughput, unbiased discovery pipeline centered on a defining catalytic product: RNA fragments with 2′,3′-cyclic phosphate and 5′-hydroxyl termini. Our key innovation is the dual-selection of both termini using RtcB ligase, which concurrently requires both cleavage signatures. This strategy dramatically enhances specificity over single-end capture methods and provides a generalizable framework for transcriptome-wide ribozyme mining.

Applying this approach, we discovered four novel self-cleaving ribozymes from the human genome. Notably, their genomic contexts suggest diverse functional and evolutionary origins. Ribozyme GS46 resides in a protein-coding exon of WDFY1, implying a potential role in post-transcriptional regulation, possibly through cleavage-mediated nonsense-mediated decay of its host mRNA. In contrast, GS59 and GS69 are embedded within repetitive elements—an intronic DNA repeat in PLD5 and an LTR retrotransposon, respectively—hinting at mobilization via genomic dynamics. The antisense localization of GS76 within an SYNJ2BP intron suggests possible cis-regulation of the sense transcript through mechanisms like RNA interference. Truncation analyses defined minimal functional cores (~70–100 nt), and computational modeling indicates these adopt novel structural folds, pointing to potentially new ribozyme classes.

Our method represents a significant advance in discovery efficiency. From 96 high-confidence candidates, we validated six active sequences—including two known and four novel ribozymes—substantially outperforming previous motif-, similarity-based and biochemical enrichment-based searches. However, the presence of false positives underscores the necessity for rigorous experimental validation. These signals may originate from trans-cleavage events, non-specific enzymatic reactions, or other RNA processing activities; their resolution will be crucial for future pipeline optimization.

A technical limitation of our RtcB-based capture is variable detection sensitivity across ribozyme classes, as evidenced by the weak signal for hovlinc and the absence of signal for CPEB3 in 5OH-seq data. Stable 3′-terminal secondary structures or post-ligation cleavage may impair adapter ligation efficiency, suggesting areas for methodological refinement.

In summary, this work significantly expands the catalog of human self-cleaving ribozymes and provides a powerful, adaptable strategy for their discovery. The unique genomic associations and putative novel structures of the identified ribozymes open new avenues for exploring RNA catalysis in eukaryotic regulation. Future efforts to determine their tertiary structures, precise mechanisms, and biological functions will deepen our understanding of the functional complexity of the human RNA world.

Methods

Experimental process for ribozyme identification

HeLa genomic DNA libraries were constructed using the tagmentation method, which simultaneously fragments genomic DNA (NEB, no. N4006S) and incorporates adapter sequences via the Tn5 transposase. The process was performed using the Nextera XT DNA Library Preparation Kit (Illumina), following the manufacturer's guidelines. PCR amplification of the tagmentation product was carried out using primers T7p-s7 and T7t-s5. These primers were designed to anneal to the Nextera s7 and s5 adapter regions, respectively, and incorporated the T7 promoter sequence at the 5′ ends of the amplicons. Separation of the DNA library fragments (500-1000 bp) was achieved by agarose gel electrophoresis and subsequent gel purification.

Purified DNA library was then used as the template for in vitro transcription by incubation with 100 U of T7 RNA Polymerase (NEB, no. M0251L) in 1×T7 RNA polymerase buffer (40 mM Tris-HCl pH 7.9, 6 mM $MgCl_2$, 1 mM DTT, 2 mM spermidine) with 2 mM NTP mix (NEB, no. N0466S) and 20 U Murine RNase Inhibitor (NEB, no. M0314S) in a 30 µl reaction volume for 5 h at 37 °C. DNA templates were removed by incubation with 2 U of DNase I (NEB, no. M0303S) and 1× DNase I reaction buffer in a 100 µl volume at 37 °C for 15 min, followed by purification with RNA Clean & Concentrator-5 kit (Zymo Research, no. R1016).

2.5 µg purified RNA were incubated with 100 pmol linker (rM13F_3desBio for 3P-seq, rM13R_5desBio_3P for 5OH-seq) in 1× RtcB reaction buffer with 22.5 pmol RtcB ligase (NEB, no. M0458S), 0.1 mM GTP and 1 mM $MnCl_2$ in 30 µl at 37 °C for 1 h. Ligation products were purified with Sera-Mag Streptavidin-Coated Magnetic Beads (Cytiva, no. 30152103010150) and eluted in 20 µl TNB buffer (20 mM Tris-HCl pH 7.5, 150 mM NaCl, 4 mM biotin). Before reverse transcription, ligation products were exchanged into H_2O by column purification.

Ligation products were reversed transcribed using primer (RT1_m13f_adp1 for 3P-seq, RT1_Tn5ME_adp1 for 5OH-seq) and the ProtoScript II Reverse Transcriptase (NEB, no. M0368S). Ligation product was first denatured at 65 °C for 5 min in a 19.5 µl volume containing 1 µl of 100 µM primer and 1.5 µl of 10 mM dNTP mix, followed by rapid snap-cooling on ice for 2 min. Then, 6 µl of 5 × ProtoScript II Reverse Transcriptase buffer, 3 µl of 0.1 M DTT, 0.5 µl of Murine RNase Inhibitor (NEB, no. M0314S) and 1 µl of ProtoScript II Reverse Transcriptase were added and incubated at 42 °C for 1 h and then 65 °C for 20 minutes to inactivate the enzyme, followed by enrichment with Sera-Mag Streptavidin-Coated Magnetic Beads (Cytiva, no. 30152103010150) and clean-up by using RNA Clean & Concentrator-5 kit (Zymo Research, no. R1016).

Purified cDNA was amplified by touch-up PCR using primers (P7R2_Tn5ME/P5R1_adp1 for 3P-seq, P7R2_m13r/P5R1_adp1 for 5OH-seq) with Q5 Hot Start High-Fidelity DNA Polymerase (NEB, no. M0493S) to construct the high-throughput sequencing library. The PCR products were purified with homemade magnetic beads mix, and sequenced on an Illumina HiSeq X sequencer with 20% PhiX control by Novogene Technology Co., Ltd.

High-throughput sequencing data analysis

A standard bioinformatic pipeline was implemented to process the sequencing data. Briefly, adapter sequences were trimmed from the raw reads using Cutadapt[12]. To ensure stringent removal of the

short adapter sequences flanking our library constructs, the Cutadapt-processed reads were further filtered with a custom script. This step involved removing the last 5 nucleotides from all reads longer than 113 nt. The resulting high-quality paired-end reads were then aligned to the UCSC human reference genome (hg38) using Bowtie2[13] in the --very-sensitive mode. The subsequent SAM file was processed to retain only uniquely mapped, properly paired reads with a MAPQ score of 30 or higher. Finally, the alignment data for each read was converted into cleavage site information, and all potential sites were statistically tallied and ranked.

Cleavage sites that appeared only once were filtered out, retaining only those sites with a read count of at least two. Noise reduction was then performed using a custom script. This method functions by analyzing the signal intensity ratios within sequencing reads of a defined window size. All potential cleavage sites within a defined genomic window (50nt) are removed if more than one site is present and the signal intensity ratio between them is less than 10. Each potential cleavage site was subsequently assigned a score calculated as (N_5OH + 1) × (N_3P + 1), where N_5OH and N_3P represent the number of occurrences of the site in the 3P-seq sequencing library and the 5OH-seq sequencing library, respectively. The top 96 ranked sequences by score, were selected as the final candidate pool for experimental validation.

Experimental validation of candidate sequences

The top 96 candidate sequences, each flanked with the T7 promoter and unique index sequences, were synthesized as an oligo pool by IDT (Integrated DNA Technologies). The oligo pool was subsequently PCR-amplified with specific forward and reverse primers to obtain the target genes. PCR products were cloned into the pUC57 vector through homologous ends encoding the T7 promoter and a NotI site by using the HiFi DNA Assembly Master Mix (NEB, no. E2621L). After chemical transformation, single colonies containing the candidate sequences were picked for downstream validation. The candidate DNA templates were transcribed in vitro for 5 hours using T7 RNA Polymerase (NEB, no. M0251L). The transcription products were digested with DNase I (NEB, no. M0303S) and denatured at 70°C for 5 min after adding an equal volume of 2 × RNA Loading Dye (NEB, no. B0363S) for subsequent denaturing PAGE analysis.

Experimental validation of ribozyme variants

All truncated variants were generated via in vitro recombination using a Seamless Cloning kit (Beyotime, no. D7010M). Specifically, each variant was amplified with a pair of primers containing

homologous arms flanking the T7 promoter and a NotI restriction site, and then cloned into the pUC57 vector through homologous recombination. Variants containing SNPs were constructed by site-directed mutagenesis using primers encoding the desired point mutations. The entire plasmid was amplified with these mutagenic primers, and the parental DNA template was digested with DpnI (NEB, no. R0176S). The digested product was subsequently transformed into E. coli DH5α cells. All constructed variants were verified by Sanger sequencing. Plasmid DNA from sequence-verified clones was linearized by NotI (NEB, no. R3189S) digestion and served as the template for in vitro transcription. Transcription reactions were performed using the T7 RNA Polymerase (Beyotime, no. R7012L), following the manufacturer's protocol. After transcription, the DNA template was removed by DNase I (NEB, no. M0303S) treatment. The RNA products were then incubated in 1× TKMS buffer (50 mM Tris-HCl pH 8.0, 25mM KCl, 10 mM $MgCl_2$, 2 mM spermidine) at 37°C for 2 hours. Finally, the RNA was purified and analyzed by 10% denaturing polyacrylamide gel electrophoresis (PAGE), followed by staining with SYBR Gold nucleic acid gel stain (Thermo Fisher, no. S11494).

Reference

1. Cech, T. R. Evolution of biological catalysis: ribozyme to RNP enzyme. *Cold Spring Harb Symp Quant Biol* **74**, 11–16 (2009).

2. Jimenez, R. M., Polanco, J. A. & Lupták, A. Chemistry and biology of self-cleaving ribozymes. *Trends Biochem Sci* **40**, 648–661 (2015).

3. Roth, A. *et al.* A widespread self-cleaving ribozyme class is revealed by bioinformatics. *Nat Chem Biol* **10**, 56–60 (2014).

4. Weinberg, Z. *et al.* New classes of self-cleaving ribozymes revealed by comparative genomics analysis. *Nat. Chem. Biol.* **11**, 606–610 (2015).

5. de la Peña, M. & García-Robles, I. Intronic hammerhead ribozymes are ultraconserved in the human genome. *EMBO Rep.* **11**, 711–716 (2010).

6. Salehi-Ashtiani, K., Lupták, A., Litovchick, A. & Szostak, J. W. A genomewide search for ribozymes reveals an HDV-like sequence in the human CPEB3 gene. *Science* **313**, 1788–1792 (2006).

7. Zhang, Z. *et al.* Minimal twister sister-like self-cleaving ribozymes in the human genome revealed by deep mutational scanning. *Elife* **12**, RP90254 (2024).

8. Chen, Y. *et al.* Hovlinc is a recently evolved class of ribozyme found in human lncRNA. *Nat Chem Biol* https://doi.org/10.1038/s41589-021-00763-0 (2021) doi:10.1038/s41589-021-00763-0.

9. Tanaka, N., Chakravarty, A. K., Maughan, B. & Shuman, S. Novel mechanism of RNA repair by RtcB via sequential 2',3'-cyclic phosphodiesterase and 3'-Phosphate/5'-hydroxyl ligation reactions. *J. Biol. Chem.* **286**, 43134–43143 (2011).

10. Ramírez, F. *et al.* deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic Acids Res* **44**, W160-165 (2016).

11. Singh, J., Hanson, J., Paliwal, K. & Zhou, Y. RNA secondary structure prediction using an ensemble of two-dimensional deep neural networks and transfer learning. *Nat Commun* **10**, 5407 (2019).

12. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* **17**, 10–12 (2011).

13. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nature Methods* **9**, 357–359 (2012).

Genome-Wide Dual-Selection Unveils Novel Self-Cleaving Ribozymes in the Human Genome

Zhe Zhang[1], Jian Zhan[1,2,3*] and Yaoqi Zhou[1*]

[1]Institute for Systems and Physical Biology, Shenzhen Bay Laboratory, Shenzhen 518107, China
[2]Ribopeutic (Shenzhen) Co., Ltd., Futian, Shenzhen, Guangdong Province, 518000, China
[3]Ribopeutic Inc., Qiantang, Hangzhou, Zhejiang Province, 310018, China
[*] Co-corresponding authors (zhanjian@szbl.ac.cn and zhouyq@szbl.ac.cn)

Abstract

The landscape of catalytic RNA in complex eukaryotes remains poorly charted. Although self-cleaving ribozymes are widespread in microbial and viral genomes, their existence and functional roles in humans remain limited. Here, we introduce a generalizable, genome-wide discovery platform that integrates high-throughput signals from two complementary adapter ligation assays—3P-seq and 5OH-seq—to specifically capture RNA fragments bearing cleavage signatures (2′,3′-cyclic phosphate/3′-phosphate and 5′-hydroxyl termini). By applying a dedicated computational scoring algorithm to human genomic data, we systematically identified four previously unrecognized self-cleaving ribozymes. These ribozymes localize to diverse genomic features: an exon of WDFY1, an intron of PLD5 embedded within a repetitive element, an LTR retrotransposon, and the antisense strand of an SYNJ2BP intron. Truncation analyses defined minimal functional cores of ~70–100 nucleotides, and computational modeling suggests that they adopt novel structural architectures. Our work establishes a powerful strategy for transcriptome-wide mining of catalytic RNA, expands the catalogue of human ribozymes from 6 to 10, and reveals new layers of functional complexity in the human RNA world.

Main

Self-cleaving Ribozymes are an important class of non-coding RNAs with catalytic functions[1] that catalyze the site-specific self-cleavage of phosphodiester backbone[2]. Recent advancements in computational and high-throughput techniques have significantly accelerated the pace of ribozyme discovery[3,4]. However, the vast majority of self-cleaving ribozymes identified to date originate from lower organisms. Only 6 self-cleavage ribozymes are found in human so far (hammerhead ribozymes (HH9, HH10) located with introns[5], the HDV-like cytoplasmic polyadenylation element-binding protein 3 (CPEB3) ribozyme[6], 2 twister sister-like (TS-like) ribozymes[7] (LINE-1 and OR4K15), and the hovlinc ribozyme found in very long intergenic non-coding RNAs (vlincRNAs)[8]. Although these self-cleaving ribozymes lack structural similarity, they all accomplish site-specific

cleavage through a nucleophilic attack by a 2′-oxygen on the adjacent phosphodiester bond. This reaction produces cleavage termini characterized by a 2′,3′-cyclic phosphate group and a 5′-hydroxyl group, respectively.

Previously, the Kapranov group utilized the RppH 5′ pyrophosphohydrolase and the XRN-1 5′-3′ exoribonuclease to degrade the other transcriptional products, except for the RNA fragments containing the 5′-hydroxyl. This led to the discovery of the hovlinc ribozyme among 28 candidates. Although the low discovery rate of self-cleaving ribozymes suggests the rarity in complex organisms, the question of whether other self-cleaving ribozymes exist in the human genome merits further study.

To probe the genome for potential self-cleaving ribozymes, we developed a strategy that leverages the unique specificity of RtcB ligase, an enzyme that joins a 5′-hydroxyl terminus to a 3′-phosphate or 2′,3′-cyclic phosphate terminus [9]. This was achieved by utilizing RtcB ligase in conjunction with specifically designed adapter RNAs to capture RNA molecules bearing these characteristic ends generated from self-cleavage. Among 96 candidate sequences subjected to experimental verification, six were confirmed to exhibit self-cleaving activity. Two of these active sequences had been documented previously, whereas the other four represent novel self-cleaving ribozymes. It is noteworthy that one self-cleaving ribozyme was identified within an exonic region of the WDFY1 (WD repeat and FYVE domain containing 1) protein, suggesting that it may play a direct role in regulating protein expression. To the best of our knowledge, this is the first report of a self-cleaving ribozyme residing in a protein-coding exon, and thus, significantly expands our understanding of the functional repertoire and mechanistic diversity of self-cleaving ribozymes.

Results

Genome-wide discovery of self-cleaving ribozymes

Genomic libraries were prepared using the Illumina tagmentation kit, which employs a Tn5 transposase to simultaneously fragment and tagmate DNA, streamlining the library construction process. Briefly, the tagmentation reaction fragmented genomic DNA and ligated Nextera s7 and Nextera s5 adapter sequences to both ends of the fragments in a single step. Then, the termini of the genomic library fragments were appended with T7 promoter through PCR amplification, facilitating

downstream in vitro transcription (Fig. 1a). These RNA transcripts along with self-cleavage products obtained from the genomic library offers the distinct advantage of being unaffected by intracellular proteases. Unlike prior methods[8] that solely enriched for 5′-hydroxyl ends from self-cleavage, our approach specifically captures cleavage products containing either 2′,3′-cyclic phosphate or 5′-hydroxyl termini from self-cleavage. The 3P-seq is designed to isolate RNA fragments possessing 2′,3′-cyclic phosphate or 3′-phosphate ends. These 3′-ends were joined with a synthetic RNA linker that features a 5′-hydroxyl group by taking advantage of the unique ligation activity of RtcB ligase. Meanwhile, the 5OH-seq captures RNAs with 5′-hydroxyl termini by employing RtcB ligates these 5′-hydroxyl ends to a separate RNA linker that itself carries a 3′-phosphate group. A desthiobiotin modification was strategically engineered at the distal ends of both RNA linkers with a dual purpose: first, to minimize the formation of off-target ligation products, and second, to enable highly specific capture and enrichment by using streptavidin-based purification methods. The enriched products were subsequently processed into sequencing libraries for high-throughput analysis.

To minimize false positives, we employed paired-end sequencing and implemented a stringent quality filter based on the Mapping Quality (MAPQ) score from the SAM files, retaining only alignments with a MAPQ $\geqslant$ 30, which is widely regarded as a threshold for high-confidence mappings. The precise cleavage sites were inferred from the alignment boundaries, corresponding to the junctions between the inserted fragments and the adapter sequences. The distribution of these putative cleavage sites was then quantified by counting the supporting reads. During data processing, we observed that a substantial number of cleavage sites remained clustered within specific genomic regions. Given that self-cleaving ribozymes typically catalyze hydrolysis at unique sites, we rationalized that sites supported by an exceptionally low number of reads within these clusters were likely artifacts, potentially resulting from random RNA degradation. Consequently, a sophisticated denoising step was employed, which eliminated sites based on an intra-window signal intensity ratio. Furthermore, to enhance the reliability of our findings, we integrated the results from two independent experiments—3P-seq and 5OH-seq. A composite scoring system was designed to prioritize cleavage sites detected in both assays, under the premise that such "consensus" sites exhibit a lower probability of being false positives.

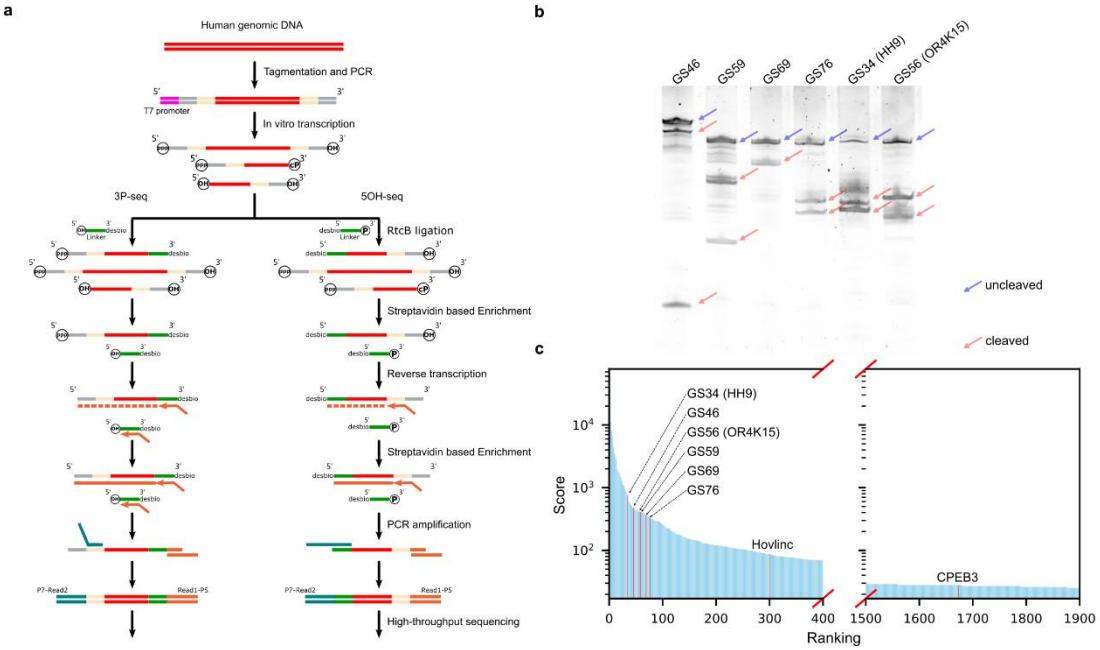

Fig. 1: Genome-scale discovery of self-cleaving ribozymes in human genome. a, The discovery flowchart: To profile RNA cleavage products, we generated a protease-resistant genomic RNA library through in vitro transcription. RtcB ligase and terminally modified RNA linkers were utilized to selectively capture and distinguish fragments bearing 2′,3′-cyclic phosphate or 5′-OH ends, followed by streptavidin bead enrichment and sequencing library construction. b, Urea PAGE analysis of the six experimentally validated ribozyme candidates from the top 96 hits. Hit nomenclature reflects its ranking (e.g., GS46 corresponds to the hit ranked 46th). c, Scoring distribution of the top 1,900 hits. Hits corresponding to known ribozymes and novel ribozymes that have been validated through experiments are explicitly marked in red.

Self-cleaving ribozymes identified in HeLa genome

Applying this integrated approach, we successfully identified four out of the five known self-cleaving ribozymes, with HH9 and OR4K15 ranking within the top 96 hits (Fig. 1b). However, the hovlinc ribozyme, discovered more recently, was ranked 301st (Fig. 1c), primarily due to its low detection count (2 reads) in the 5OH-seq assay. The CPEB3 ribozyme was ranked 1,673rd (Fig. 1c) and was not detected in the 5′OH-seq assay.

To gain some understanding of low priority in hovlinc and CPEB3 ribozymes, we generated normalized coverage files from the BAM files and visualized the signal intensity across the targeted gene regions by using the deepTools suite[10]. Analysis of the signal profiles across identified

ribozymes shows a spectrum: from those with a pronounced signal asymmetry (Fig. 2a, 2c, 2f, 2g and 2h) to those with relatively balanced signals at both ends (Fig. 2b, 2d and 2e). This variation is principally dictated by the unique sequence and the consequent secondary/tertiary structure of each ribozyme, which directly affects the mapping efficacy of the RNA linkers during the library preparation. These inherent structural differences are a major source of the significant variation in observed sequencing depth. For hovlinc and CPEB3 ribozymes, their characteristically stable 3′ terminal structures likely impede efficient adapter ligation, resulting in negligible 3′ end capture signal and thus a substantially lower ranking in our analysis, highlighting a specific limitation in detecting self-cleaving ribozymes with such structural features.

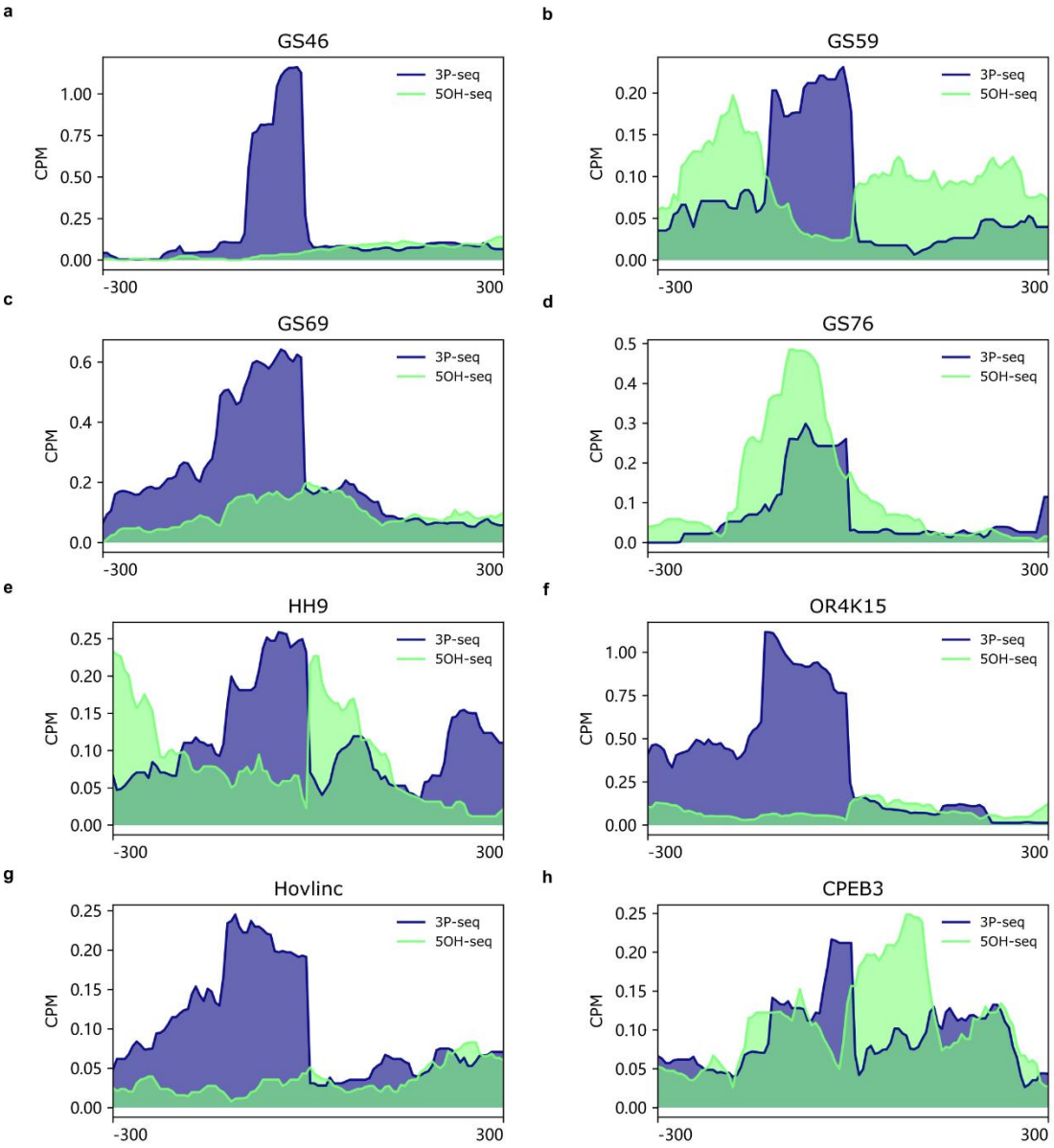

Fig. 2: Normalized signal from the two genome-wide selection assays. The y-axis represents the counts per million (CPM) of mapped reads, while the x-axis corresponds to a 600-nucleotide (nt) genomic window centered on the cleavage site. a–d illustrate four novel ribozymes identified in this study. e and f depict two previously reported ribozymes (HH9 ranked 34th, OR4K15 ranked 56th) that were confirmed among the top 96 validated sequences in our screening. In contrast, g and h show two known ribozymes that ranked lower in our results, indicating their relatively lower abundance or activity under the experimental conditions.

New ribozymes with potential novel function

We characterized four novel self-cleaving ribozymes (GS46, GS59, GS69, GS76), each occupying a distinct genomic locus that suggests unique regulatory potentials. Particularly notable is GS46, which was located within the eighth exon of the protein-coding gene WDFY1 (WD repeat and FYVE domain containing 1); its self-cleavage activity is thus supposed to truncate the host mRNA, potentially triggering nonsense-mediated decay and leading to reduced expression of the WDFY1 protein. Furthermore, GS59 and GS69 were identified within repetitive genomic architectures: GS59 was situated in the first intron of PLD5 (Phospholipase D Family Member 5) and embedded within a DNA repeat element, while GS69 was found within an intron of a novel transcript that itself resides in a Long Terminal Repeat (LTR) retrotransposon. Additionally, GS76 was mapped to the reverse (antisense) strand of the first intron of the SYNJ2BP (Synaptojanin 2 Binding Protein) gene, an intriguing location that raises the possibility of its involvement in regulating the sense transcript.

To elucidate the structural basis of the self-cleaving activity, we conducted a preliminary investigation of the four newly identified ribozymes. It is important to note that the RNA fragments captured in our sequencing libraries do not necessarily represent the minimal structural module required for catalytic activity as we demonstrated previously[7]. Therefore, to delineate the core functional domain of each ribozyme, we systematically generated a series of truncated variants. Given that the 3P-seq assay yielded stronger signals for most of these ribozymes (Fig. 2), indicating a more defined or stable 3′ end post-cleavage, the truncation strategy primarily involved progressive shortening from the 3′ terminus. The catalytic competence of these truncated constructs was then assessed by analyzing their self-cleavage efficiency via polyacrylamide gel

electrophoresis (PAGE). This empirical approach allowed us to identify the minimal functional core for each ribozyme, with optimized lengths of approximately 70 nt, 100 nt, 80 nt, and 80 nt, respectively (Fig. 3a). Subsequent computational prediction of the secondary and tertiary structures for these minimized variants suggested that each adopts a distinct and previously uncharacterized structural fold (Fig. 3b), potentially representing novel architectural classes among self-cleaving ribozymes.

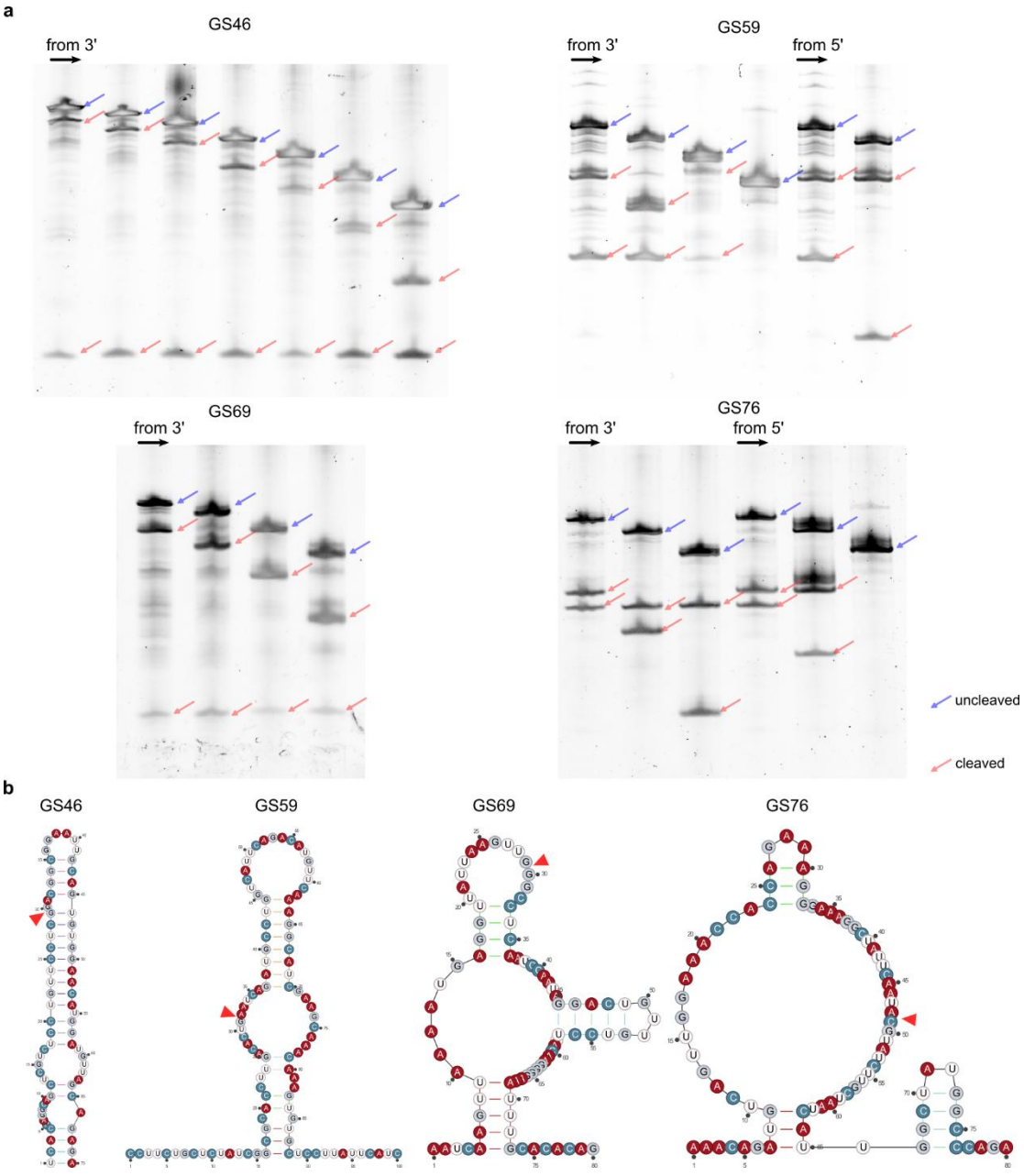

Fig. 3: Truncations and predicted secondary structures of the 4 new ribozymes. a, PAGE analysis of the truncated variants of the 4 new ribozymes. A series of truncated variants was generated from the original sequence for activity assay. The 3′ end (or 5′ end) was anchored, and sequential 20-nt

deletions were made from the opposite terminus. This process continued until the whole fragment reached a minimum length of 100 nt, which was determined to be optimal for efficient molecular cloning. b, Predicted secondary structure of the minimal truncated ribozymes. The secondary structures of the RNA sequences were predicted by employing the computational tool SPOT-RNA[11]. The red triangle indicates the cleavage site.

Discussion

Self-cleaving ribozymes are ubiquitous in lower organisms, yet their representation in complex eukaryotic genomes remains enigmatic. Although prior studies have identified a handful of examples in humans—including the HDV-like CPEB3 ribozyme[6], two hammerhead ribozymes[5], and the hovlinc ribozyme[8]—these discoveries relied heavily on known sequence motifs， specific biochemical properties or genome-scale search according to the discovery of hovlinc ribozyme[8]. This raises a fundamental question: is the scarcity of known human ribozymes biological reality, or a reflection of methodological constraints?

To address this, we developed a high-throughput, unbiased discovery pipeline centered on a defining catalytic product: RNA fragments with 2′,3′-cyclic phosphate and 5′-hydroxyl termini. Our key innovation is the dual-selection of both termini using RtcB ligase, which concurrently requires both cleavage signatures. This strategy dramatically enhances specificity over single-end capture methods and provides a generalizable framework for transcriptome-wide ribozyme mining.

Applying this approach, we discovered four novel self-cleaving ribozymes from the human genome. Notably, their genomic contexts suggest diverse functional and evolutionary origins. Ribozyme GS46 resides in a protein-coding exon of WDFY1, implying a potential role in post-transcriptional regulation, possibly through cleavage-mediated nonsense-mediated decay of its host mRNA. In contrast, GS59 and GS69 are embedded within repetitive elements—an intronic DNA repeat in PLD5 and an LTR retrotransposon, respectively—hinting at mobilization via genomic dynamics. The antisense localization of GS76 within an SYNJ2BP intron suggests possible cis-regulation of the sense transcript through mechanisms like RNA interference. Truncation analyses defined minimal functional cores (~70–100 nt), and computational modeling indicates these adopt novel structural folds, pointing to potentially new ribozyme classes.

Our method represents a significant advance in discovery efficiency. From 96 high-confidence candidates, we validated six active sequences—including two known and four novel ribozymes—substantially outperforming previous motif-, similarity-based and biochemical enrichment-based searches. However, the presence of false positives underscores the necessity for rigorous experimental validation. These signals may originate from trans-cleavage events, non-specific enzymatic reactions, or other RNA processing activities; their resolution will be crucial for future pipeline optimization.

A technical limitation of our RtcB-based capture is variable detection sensitivity across ribozyme classes, as evidenced by the weak signal for hovlinc and the absence of signal for CPEB3 in 5OH-seq data. Stable 3′-terminal secondary structures or post-ligation cleavage may impair adapter ligation efficiency, suggesting areas for methodological refinement.

In summary, this work significantly expands the catalog of human self-cleaving ribozymes and provides a powerful, adaptable strategy for their discovery. The unique genomic associations and putative novel structures of the identified ribozymes open new avenues for exploring RNA catalysis in eukaryotic regulation. Future efforts to determine their tertiary structures, precise mechanisms, and biological functions will deepen our understanding of the functional complexity of the human RNA world.

Methods

Experimental process for ribozyme identification

HeLa genomic DNA libraries were constructed using the tagmentation method, which simultaneously fragments genomic DNA (NEB, no. N4006S) and incorporates adapter sequences via the Tn5 transposase. The process was performed using the Nextera XT DNA Library Preparation Kit (Illumina), following the manufacturer's guidelines. PCR amplification of the tagmentation product was carried out using primers T7p-s7 and T7t-s5. These primers were designed to anneal to the Nextera s7 and s5 adapter regions, respectively, and incorporated the T7 promoter sequence at the 5′ ends of the amplicons. Separation of the DNA library fragments (500-1000 bp) was achieved by agarose gel electrophoresis and subsequent gel purification.

Purified DNA library was then used as the template for in vitro transcription by incubation with 100 U of T7 RNA Polymerase (NEB, no. M0251L) in 1×T7 RNA polymerase buffer (40 mM Tris-HCl pH 7.9, 6 mM $MgCl_2$, 1 mM DTT, 2 mM spermidine) with 2 mM NTP mix (NEB, no. N0466S) and 20 U Murine RNase Inhibitor (NEB, no. M0314S) in a 30 µl reaction volume for 5 h at 37 °C. DNA templates were removed by incubation with 2 U of DNase I (NEB, no. M0303S) and 1× DNase I reaction buffer in a 100 µl volume at 37 °C for 15 min, followed by purification with RNA Clean & Concentrator-5 kit (Zymo Research, no. R1016).

2.5 µg purified RNA were incubated with 100 pmol linker (rM13F_3desBio for 3P-seq, rM13R_5desBio_3P for 5OH-seq) in 1× RtcB reaction buffer with 22.5 pmol RtcB ligase (NEB, no. M0458S), 0.1 mM GTP and 1 mM $MnCl_2$ in 30 µl at 37 °C for 1 h. Ligation products were purified with Sera-Mag Streptavidin-Coated Magnetic Beads (Cytiva, no. 30152103010150) and eluted in 20 µl TNB buffer (20 mM Tris-HCl pH 7.5, 150 mM NaCl, 4 mM biotin). Before reverse transcription, ligation products were exchanged into H_2O by column purification.

Ligation products were reversed transcribed using primer (RT1_m13f_adp1 for 3P-seq, RT1_Tn5ME_adp1 for 5OH-seq) and the ProtoScript II Reverse Transcriptase (NEB, no. M0368S). Ligation product was first denatured at 65 °C for 5 min in a 19.5 µl volume containing 1 µl of 100 µM primer and 1.5 µl of 10 mM dNTP mix, followed by rapid snap-cooling on ice for 2 min. Then, 6 µl of 5 × ProtoScript II Reverse Transcriptase buffer, 3 µl of 0.1 M DTT, 0.5 µl of Murine RNase Inhibitor (NEB, no. M0314S) and 1 µl of ProtoScript II Reverse Transcriptase were added and incubated at 42 °C for 1 h and then 65 °C for 20 minutes to inactivate the enzyme, followed by enrichment with Sera-Mag Streptavidin-Coated Magnetic Beads (Cytiva, no. 30152103010150) and clean-up by using RNA Clean & Concentrator-5 kit (Zymo Research, no. R1016).

Purified cDNA was amplified by touch-up PCR using primers (P7R2_Tn5ME/P5R1_adp1 for 3P-seq, P7R2_m13r/P5R1_adp1 for 5OH-seq) with Q5 Hot Start High-Fidelity DNA Polymerase (NEB, no. M0493S) to construct the high-throughput sequencing library. The PCR products were purified with homemade magnetic beads mix, and sequenced on an Illumina HiSeq X sequencer with 20% PhiX control by Novogene Technology Co., Ltd.

High-throughput sequencing data analysis

A standard bioinformatic pipeline was implemented to process the sequencing data. Briefly, adapter sequences were trimmed from the raw reads using Cutadapt[12]. To ensure stringent removal of the

short adapter sequences flanking our library constructs, the Cutadapt-processed reads were further filtered with a custom script. This step involved removing the last 5 nucleotides from all reads longer than 113 nt. The resulting high-quality paired-end reads were then aligned to the UCSC human reference genome (hg38) using Bowtie2[13] in the --very-sensitive mode. The subsequent SAM file was processed to retain only uniquely mapped, properly paired reads with a MAPQ score of 30 or higher. Finally, the alignment data for each read was converted into cleavage site information, and all potential sites were statistically tallied and ranked.

Cleavage sites that appeared only once were filtered out, retaining only those sites with a read count of at least two. Noise reduction was then performed using a custom script. This method functions by analyzing the signal intensity ratios within sequencing reads of a defined window size. All potential cleavage sites within a defined genomic window (50nt) are removed if more than one site is present and the signal intensity ratio between them is less than 10. Each potential cleavage site was subsequently assigned a score calculated as (N_5OH + 1) × (N_3P + 1), where N_5OH and N_3P represent the number of occurrences of the site in the 3P-seq sequencing library and the 5OH-seq sequencing library, respectively. The top 96 ranked sequences by score, were selected as the final candidate pool for experimental validation.

Experimental validation of candidate sequences

The top 96 candidate sequences, each flanked with the T7 promoter and unique index sequences, were synthesized as an oligo pool by IDT (Integrated DNA Technologies). The oligo pool was subsequently PCR-amplified with specific forward and reverse primers to obtain the target genes. PCR products were cloned into the pUC57 vector through homologous ends encoding the T7 promoter and a NotI site by using the HiFi DNA Assembly Master Mix (NEB, no. E2621L). After chemical transformation, single colonies containing the candidate sequences were picked for downstream validation. The candidate DNA templates were transcribed in vitro for 5 hours using T7 RNA Polymerase (NEB, no. M0251L). The transcription products were digested with DNase I (NEB, no. M0303S) and denatured at 70°C for 5 min after adding an equal volume of 2 × RNA Loading Dye (NEB, no. B0363S) for subsequent denaturing PAGE analysis.

Experimental validation of ribozyme variants

All truncated variants were generated via in vitro recombination using a Seamless Cloning kit (Beyotime, no. D7010M). Specifically, each variant was amplified with a pair of primers containing

homologous arms flanking the T7 promoter and a NotI restriction site, and then cloned into the pUC57 vector through homologous recombination. Variants containing SNPs were constructed by site-directed mutagenesis using primers encoding the desired point mutations. The entire plasmid was amplified with these mutagenic primers, and the parental DNA template was digested with DpnI (NEB, no. R0176S). The digested product was subsequently transformed into E. coli DH5α cells. All constructed variants were verified by Sanger sequencing. Plasmid DNA from sequence-verified clones was linearized by NotI (NEB, no. R3189S) digestion and served as the template for in vitro transcription. Transcription reactions were performed using the T7 RNA Polymerase (Beyotime, no. R7012L), following the manufacturer's protocol. After transcription, the DNA template was removed by DNase I (NEB, no. M0303S) treatment. The RNA products were then incubated in 1× TKMS buffer (50 mM Tris-HCl pH 8.0, 25mM KCl, 10 mM $MgCl_2$, 2 mM spermidine) at 37°C for 2 hours. Finally, the RNA was purified and analyzed by 10% denaturing polyacrylamide gel electrophoresis (PAGE), followed by staining with SYBR Gold nucleic acid gel stain (Thermo Fisher, no. S11494).

Reference

1. Cech, T. R. Evolution of biological catalysis: ribozyme to RNP enzyme. *Cold Spring Harb Symp Quant Biol* **74**, 11–16 (2009).

2. Jimenez, R. M., Polanco, J. A. & Lupták, A. Chemistry and biology of self-cleaving ribozymes. *Trends Biochem Sci* **40**, 648–661 (2015).

3. Roth, A. *et al.* A widespread self-cleaving ribozyme class is revealed by bioinformatics. *Nat Chem Biol* **10**, 56–60 (2014).

4. Weinberg, Z. *et al.* New classes of self-cleaving ribozymes revealed by comparative genomics analysis. *Nat. Chem. Biol.* **11**, 606–610 (2015).

5. de la Peña, M. & García-Robles, I. Intronic hammerhead ribozymes are ultraconserved in the human genome. *EMBO Rep.* **11**, 711–716 (2010).

6. Salehi-Ashtiani, K., Lupták, A., Litovchick, A. & Szostak, J. W. A genomewide search for ribozymes reveals an HDV-like sequence in the human CPEB3 gene. *Science* **313**, 1788–1792 (2006).

7. Zhang, Z. *et al.* Minimal twister sister-like self-cleaving ribozymes in the human genome revealed by deep mutational scanning. *Elife* **12**, RP90254 (2024).

8. Chen, Y. *et al.* Hovlinc is a recently evolved class of ribozyme found in human lncRNA. *Nat Chem Biol* https://doi.org/10.1038/s41589-021-00763-0 (2021) doi:10.1038/s41589-021-00763-0.

9. Tanaka, N., Chakravarty, A. K., Maughan, B. & Shuman, S. Novel mechanism of RNA repair by RtcB via sequential 2',3'-cyclic phosphodiesterase and 3'-Phosphate/5'-hydroxyl ligation reactions. *J. Biol. Chem.* **286**, 43134–43143 (2011).

10. Ramírez, F. *et al.* deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic Acids Res* **44**, W160-165 (2016).

11. Singh, J., Hanson, J., Paliwal, K. & Zhou, Y. RNA secondary structure prediction using an ensemble of two-dimensional deep neural networks and transfer learning. *Nat Commun* **10**, 5407 (2019).

12. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* **17**, 10–12 (2011).

13. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nature Methods* **9**, 357–359 (2012).