Published online DD MM YYYY *Nucleic Acids Research, YYYY, Vol. xx, No. xx* **1–12**
doi:10.1093/nar/gkn000

DATA RESOURCES AND ANALYSES
Landscape of semi-extractable RNAs across five human cell lines

Chao Zeng [1], Takeshi Chujo [2], Tetsuro Hirose [3,4,*] and Michiaki Hamada [1,5,6,*]

[1]Faculty of Science and Engineering, Waseda University, Tokyo 1698555, Japan [2]Faculty of Life Sciences, Kumamoto University, Kumamoto 8608556, Japan [3]Graduate School of Frontier Biosciences, Osaka University, Suita 5650871, Japan [4]Institute for Open and Transdisciplinary Research Initiatives (OTRI), Osaka University, Suita 5650871, Japan [5]AIST-Waseda University Computational Bio Big-Data Open Innovation Laboratory (CBBD-OIL), National Institute of Advanced Industrial Science and Technology, Tokyo 1698555, Japan [6]Graduate School of Medicine, Nippon Medical School, Tokyo 1138602, Japan

Received YYYY-MM-DD; Revised YYYY-MM-DD; Accepted YYYY-MM-DD

ABSTRACT

Phase-separated membraneless organelles often contain RNAs that exhibit unusual semi-extractability upon the conventional RNA extraction method, and can be efficiently retrieved by needle shearing or heating during RNA extraction. Semi-extractable RNAs are promising resources for understanding RNA-centric phase separation. However, limited assessments have been performed to systematically identify and characterize semi-extractable RNAs. In this study, 1,325 semi-extractable RNAs were identified across five human cell lines, including NEAT1, TRIO, EXT1, ZCCHC7, and FTX, which exhibited stable semi-extractability. Semi-extractable RNAs tend to be distributed in the nucleolus but are dissociated from the chromatin. Long and repeat-containing semi-extractable RNAs act as hubs to provide global RNA-RNA interactions. Semi-extractable RNAs were divided into four groups based on their k-mer content. Consistently, the NEAT1 group preferred to interact with paraspeckle proteins, such as FUS and NONO, implying that RNAs in this group are potential candidates of architectural RNAs that constitute nuclear bodies.

INTRODUCTION

Liquid-liquid phase separation (LLPS) is a biological phenomenon in which macromolecules, such as proteins or nucleic acids, are spatially organized into membrane-less organelles (also called biomolecular condensates) (1). Membrane-less organelles (MLOs) usually maintain their stable structures through multivalent interactions of molecules that act in diverse biological processes ranging from macromolecular biogenesis to gene regulation (2, 3, 4). MLOs are highly dynamic structures, whose components are rapidly exchanged between other condensates and the surrounding milieu (5, 6, 7, 8, 9), implying that MLOs are sensitive to internal and external signals. LLPS provides a new framework for our understanding of human health and disease (10, 11, 12). Phase-separated MLOs that have been discovered and studied include the nucleolus, paraspeckle, nuclear speckle, Cajal body, PML nuclear body, P-body, stress granule, germ granule, and mRNP granule (3). The role of proteins in LLPS and their regulation has been the focus of attention (1, 13, 14, 15). However, based on accumulating evidence, RNAs, especially long noncoding RNAs (lncRNAs), play a crucial role in the process of phase separation (16, 17, 18, 19, 20, 21).

As a remarkable example, nuclear paraspeckle assembly transcript 1 (NEAT1) is an architectural lncRNA that mediates the assembly of paraspeckles by driving phase separation (22, 23, 24, 25). Two major isoforms are generated from the NEAT1 gene locus, and the longer isoform NEAT1_2 serves as a molecular scaffold for the formation of RNA-protein and RNA-RNA interactions (19, 26). Paraspeckles form a core-shell spheroidal structure, in which the shell contains the 5′ and 3′ regions of NEAT1_2 and some specific proteins, whereas the core consists of the middle region of NEAT1_2 and Drosophila behaviour/human splicing (DBHS) proteins (27). According to further studies, the NEAT1_2 middle region contains redundant subdomains that sequester RNA-binding proteins (RBPs), such as non-POU domain-containing octamer-binding protein (NONO) and splicing factor proline and glutamine rich (SFPQ), to initiate paraspeckle assembly (28). Note that both NONO and SFPQ are members of the DBHS family of proteins. Interestingly, when a conventional RNA extraction method using AGPC (acid guanidinium thiocyanate-phenol-chloroform) reagent such as TRIzol is employed, most of the NEAT1 is retained in the protein layer between the aqueous phase and organic phase, resulting in a low extraction level. However, after the phase-separated

*To whom correspondence should be addressed. Email: MH (mhamada@waseda.jp) or TH (hirose.tetsuro.fbs@osaka-u.ac.jp)

© YYYY The Author(s)
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/2.0/uk/) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

structures are disrupted by an improved RNA extraction through needle shearing or heating, NEAT1 is released into the aqueous solution, and its extraction level can be 20-fold higher than that obtained via the conventional. Such property of NEAT1 is termed as "semi-extractability" (29). The semi-extractability of NEAT1 strongly depended on the prion-like domain of a paraspeckle RBP, FUS, implying that extensive multivalent interactions may cause semi-extractability (22). In addition to NEAT1, several other newly detected semi-extractable RNAs were observed to form granule-like foci in a previous study (29). Accordingly, RNAs in the phase-separated structures may commonly possess semi-extractability owing to multivalent forces. The systematic identification and characterization of semi-extractable RNAs could aid in the discovery of RNAs associated with phase separated MLOs and provide insights into LLPS biology.

In this study, we developed a genome-based transcriptome assembly approach to define 1,325 semi-extractable RNAs for the first time in five human cell lines. These RNAs prefer to be transcribed from enhancer, repressed or heterochromatin regions that are clustered in the nucleolus. Long and AU-rich semi-extractable RNAs contain more repetitive sequences than expected and interact frequently with other RNAs. Semi-extractable RNAs can be broadly classified into four different groups based on their sequence composition, with the semi-extractable RNAs of the NEAT1 group preferring to bind paraspeckle RBPs (e.g., NONO and FUS), suggesting their potential role as architectural RNAs.

MATERIALS AND METHODS

RNA-seq analysis

Pair-end reads were trimmed using cutadapt (v3.5)(30) with the following parameters: -a AGATCGGAAGAGCACACGTCTGAACTCCAGTCAC -A AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTA --overlap 5 --trim-n --max-n 1 --minimum-length 50:50. For single-end reads, the adapter-removal step was skipped. First, the reads were mapped to ribosomal RNAs using STAR (v2.7.10a) (31) when multi-mapped reads were allowed. Thereafter, the unmapped reads were mapped to the genome using STAR with the following parameter: --outFilterMultimapNmax 1. Duplicate reads were removed using Picard (v2.5.0, http://broadinstitute.github.io/picard/). The human genome sequence (hg38) and basic gene annotation were downloaded from the GENCODE (v39) project (32). Only the reference chromosomes were used for subsequent analyses. Ribosomal RNAs were merged from RefSeq (release 210) (33) and Ensembl (release 105) (34).

RNA-seq, obtained using the improved RNA extraction method, was used to construct the reference transcriptome. For each sample, StringTie (v2.2.1) (35) assembles transcripts based on mapped reads with the following parameters: --rf -i -g 500 -f 0.5. Notably, mapped reads that crossed splice sites were eliminated during the assembly. The transcripts obtained from all samples were grouped by forward and reverse strands and then merged separately using StringTie based on the following parameters: --merge -g 500. Finally, the two groups of transcripts were concatenated into the reference transcriptome. A transcript was assigned with a gene name based on an overlap with the gene by more than one base. Of note, a transcript can be assigned multiple gene names. Transcript abundance (FPKM, fragments per kilobase of exon per million mapped reads) was estimated using the StringTie quantification mode (-e) with default parameters.

The UCSC genome browser (36) was used to visualize the reference transcriptome and read coverage. To visualize read coverage, the mapped reads in BAM format were indexed using Samtools (v1.14) (37) and then converted to bigWig format with bamCoverage (v3.5.1) (38) using the following parameters: --filterRNAstrand forward/reverse --scaleFactor 1/-1 -bs 1 --normalizeUsing RPKM.

Semi-extractable RNAs

For the i-th transcript (T_i) in a cell line, we estimated its expression $FPKM_i^{conv}$ and $FPKM_i^{impr}$ in the conventional RNA extraction and the improved extraction, respectively, and then calculated the average expression (E_i) and fold change (FC_i) as follows:

$$E_i = (FPKM_i^{conv} + FPKM_i^{impr})/2, \quad \textbf{(1)}$$

$$FC_i = (FPKM_i^{impr} + 1)/(FPKM_i^{conv} + 1). \quad \textbf{(2)}$$

For each cell line, semi-extractable RNAs (seRNAs) and extractable RNAs (exRNAs) were defined using the following criteria.

$$seRNAs := \{T_i \,|\, E_i \geq 1.5, FC_i \geq 1.5\}, \quad \textbf{(3)}$$

$$exRNAs := \{T_i \,|\, 1.5 \leq E_i \leq E_{NEAT1}, 0.95 \leq FC_i \leq 1.05\}. \quad \textbf{(4)}$$

where E_{NEAT1} is the average NEAT1 expression. Only a single transcript (long isoform) of NEAT1 was present in the reference transcriptome.

For subsequent meta-analysis, the set of semi-extractable RNAs (denoted as "SE") and the set of extractable RNAs (denoted as "EX") were further defined across the A10, A549, HAP1, HEK, and HeLa cell lines. Accordingly, all semi-extractable and extractable RNAs were merged separately from the five cell lines. The transcripts that overlapped between the two sets were then removed. Additionally, all annotated intron-containing transcripts were prepared as background controls (denoted as "BG").

Chromatin state

The chromatin states of HeLa cells were downloaded from the ENCODE (39) project (http://hgdownload.cse.ucsc.edu/goldenpath/hg19/encodeDCC/wgEncodeAwgSegmentation/wgEncodeAwgSegmentationChromhmmHelas3.bed.gz). These chromatin states were predicted using a trained ChromHMM (40) model based on multiple chromatin datasets, including ChIP-seq data for various histone modifications. The annotations of chromatin states that were on hg19 were remapped to hg38 using the pyliftover package (https://github.com/konstantint/pyliftover). The chromatin state prefixes were re-annotated as follows: (i) Active Promoter: Tss and TssF; (ii) Promoter Flanking:

PromF; (iii) Inactive Promoter: PromP; (iv) Candidate Strong enhancer: Enh and EnhF; (v) Candidate Weak enhancer/DNase: EnhWF, EnhW, DNaseU, DNaseD; (vi) Distal CTCF/Candidate Insulator: CtrcfO and Ctcf; (vii) Transcription associated: Gen5′, Elon, ElonW, Gen3′, Pol2, H4K20. (viii) Low activity proximal to active states: Low. (ix) Polycomb repressed: ReprD, Repr, and ReprW; and (x) Heterochromatin/Repetitive/Copy Number Variation: Quies, Art. The chromatin states were then intersected with semi-extractable and extractable RNAs using the BEDTools (41) intersect command.

Subcellular localization

APEX-seq data for HEK cells were obtained from GSE116008. APEX-seq is an RNA sequencing method coupled with direct RNA proximity labeling (42). For each cell compartment, we measured the enrichment of a transcript in that compartment (termed subcellular localization) by calculating the fold-change in the abundance of that transcript between labeled and unlabeled libraries. Accordingly, the RNA-seq reads were first subjected to adapter trimming using Trimmomatic (v0.39) (43) with the following parameters: ILLUMINACLIP:adapter.fa:2:30:4 TRAILING:20 MINLEN:36. Then the reads were uniquely mapped to the human genome using STAR, and the transcript abundance was estimated using RSEM (v1.3.3) (44). Finally, subcellular localization (log2 fold-change in transcript abundance) was calculated using an in-house script. For a transcript, a higher value of subcellular localization value indicates a higher enrichment in the corresponding cell compartment.

Minimum free energy analysis

Using a transcript, subsequences of 300 nt length were extracted from its 5′ and 3′ ends. Transcripts less than 600 nt in length were removed beforehand. These subsequences were subjected to minimum free energy (MFE) calculations using RNAfold (v2.5.0) (45) with default parameters. Generally, a lower MFE value indicates a more stable RNA structure.

RNA-chromatin interaction

In situ mapping of RNA-Genome Interactome (iMARGI) data of HEK cells were downloaded from GSM3478205. iMARGI is a DNA sequencing method based on RNA-DNA proximity ligation *in situ* inside an intact nucleus (46). The genomic coordinates of the RNA ends in the RNA-DNA interactions were extracted from the processed iMARGI data. The RNA ends were then intersected with transcripts using the BEDTools intersect command. To measure the extent to which a transcript interacts with chromatin, the fraction of transcript regions covered by iMARGI RNA ends was calculated. This fraction ranged from 0 to 1, with a higher fraction suggesting a more frequent interaction between the transcript and chromatin.

RNA-RNA interaction

RNA interaction hubs (termed "hub RNAs") were derived from a previous study (47), in which RNA-RNA interactions were quantified by the RNA *in situ* conformation sequencing (RIC-seq), a technique based on crosslinking, proximity ligation, and sequencing. Hub RNAs exhibited stronger trans-interactions than other RNAs.

Repeat density

The genomic coordinates of the repeat sequences were extracted using RepeatMasker (hg38, repeat library 20140131; https://www.repeatmasker.org/species/hg.html). For a transcript, BEDtools was used for intersection with repeat sequences. The fraction of the transcript that overlapped with repeat sequences, termed repeat density, was then calculated using an in-house script.

Sequence motif analysis

Human RBP-binding sequence motifs (position weight matrix format) were downloaded from the CISBP-RNA database (http://cisbp-rna.ccbr.utoronto.ca; accessed on March 12, 2022). For a transcript, FIMO (v5.4.1) (48) scanned RBP-binding sites based on the above motifs using the following parameters: --norc --thresh 0.01 --motif-pseudo 0.1 --max-stored-scores 100000000. Given a transcript, the binding preference of a certain RBP was defined as the number of binding sites of this RBP normalized by the transcript length.

K-mer analysis

Semi-extractable RNAs were functionally classified using the k-mer content-based SEEKR (49) algorithm. First, seekr_kmer_counts was used to count the frequency of k-mer occurrence with the following parameter: -k 6. Thereafter, seekr_pearson was used to calculate the similarity matrix. Finally, seekr_graph segmented the RNA sequences into different communities based on the similarity matrix described above and the following parameters: .13 -n 3 --louvain. The network graph of the semi-extractable RNAs was visualized using Gephi (v0.9) (50) with a Yifan Hu proportional layout.

Gene ontology analysis

Gene ontology (GO) analysis of the semi-extractable genes across the five cell lines was conducted using g:Profiler (version: e105_eg52_p16_e84549f) (51). Statistical domain scope: only annotated genes; significance threshold: Bonferroni correction; user threshold: 0.001.

Data availability

The conventional and semi-extractable RNA-seq of A10, A549, HAP1, and HEK cells have been deposited in the DDBJ Sequence Read Archive (DRA, https://www.ddbj.nig.ac.jp) under accession numbers DRA009793, DRA012807, DRA012808, DRA012810. Published RNA-seq of HeLa cells were retrieved from Gene Expression Omnibus (GEO, https://www.ncbi.nlm.nih.gov/geo) under accession number GSE80589.

RESULTS

Genome-based assembly of semi-extractable RNAs

To identify the semi-extractable RNAs, transcriptome assembly was first performed based on the RNA-seq data

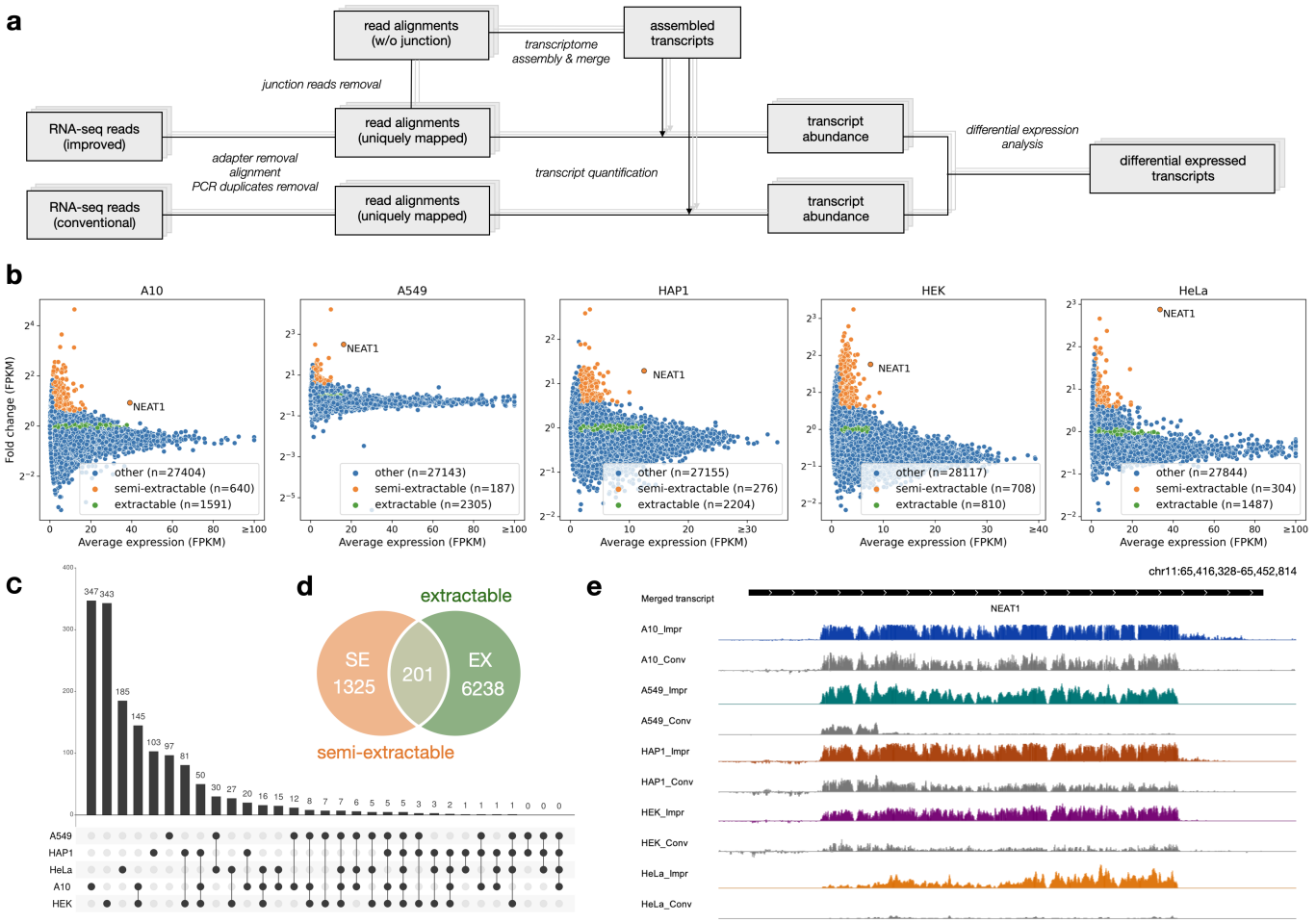

Figure 1. Identification of semi-extractable RNAs. (a) RNA-seq data analysis workflow. **(b)** Identification of semi-extractable RNAs (orange, average_expression $\geq$ 1.5 and fold_change $\geq$ 1.5) and extractable RNAs (green, 1.5 $\leq$ average_expression $\leq$ FPKM_of_NEAT1 and 0.95 $\leq$ fold_change $\leq$ 1.05) from A10, A549, HAP1, HEK, and HeLa cells. **(c)** Overlapping semi-extractable RNAs across A10, A549, HAP1, HEK, and HeLa cells. **(d)** A total of 1,325 semi-extractable RNAs (denoted as SE) and 6,238 extractable RNAs (denoted as EX) were obtained after merging from the five cells and removing the 201 overlapping RNAs. RNAs detected as semi-extractable in any of the cell lines in (b) are listed in SE. Same as EX. **(e)** NEAT1 was simultaneously detected as a semi-extractable RNA in the five cells. Impr: improved RNA extraction, Conv: conventional RNA extraction.

produced by the improved RNA extraction (Figure 1A). The rationale for this approach is based on our observation that numerous semi-extractable RNAs are not properly annotated in the existing public databases. For example, hundreds of readthrough downstream-of-gene (DoG) transcripts were discovered to be semi-extractable and reported in another study (52). Semi-extractable RNAs may be the products and intermediates of various steps (e.g., transcription, processing, and degradation) and thus contain intronic sequences or partially missing exonic sequences. Further, a semi-extractable RNA may not be available in the existing gene annotations, because it is derived from intergenic regions. We adopted a genome-based transcriptome assembly approach without reference to the gene annotations. RNA-seq reads mapped to the genome were used to construct candidate sequences for semi-extractable RNAs (Figure 1A). To improve the accuracy of transcriptome assembly, we mainly considered the following aspects: how to handle multi-mapping reads that aligned to more than one location on the genome? Do spliced isoforms arise from alternative splicing? How can we benchmark and select the appropriate assembly method? Dozens of semi-extractable RNAs validated and reported in a previous study were used to evaluate assembly performance (partial results are shown below) (29).

First, we removed the multi-mapping reads prior to the transcriptome assembly. Most of the RNA-seq reads in this study were short single-ended (36 nt), maintaining an empirically low unique mapping rate (71% on average, see Table S1). Uniquely mapped reads (referred to as uniq-reads) have a higher confidence than multi-mapping reads of ambiguous origin. For single-ended reads, multi-mapping reads tended to cause higher read coverage over regions including simple repeat and/or low-complexity sequences Figure S1, bottom). Such ambiguous regions even confounded the surrounding high-confidence regions that were supported by uniq-reads. The above situation was not alleviated by applying longer pair-ended reads (101 nt, Figure S1, upper). Accordingly, we concluded that multi-mapping reads may

lead to ambiguous transcript assembly, especially in regions containing simple repeats and/or low-complexity sequences.

The reads across splice junctions were marked to prevent them from engaging in transcriptome assembly; this is because alternative splicing leads to higher transcriptome complexity (i.e., the number of spliced isoforms of each gene). Transcript quantification based on a complex transcriptome will is challenging. In addition, the short single-ended reads used in this study might lead to the limited accuracy in detecting spliced isoforms (53). Notably, we found that most of the evaluated semi-extractable RNAs retained their intronic sequences, implying that the semi-extractable RNAs had not yet been spliced. This observation is consistent with that in a previous study (29).

We compared several popular assembly strategies to obtain an appropriate transcriptome, including two transcriptome assemblers (Cufflinks and StringTie) and two peak callers (MACS and Homer). Additionally, we parallelly compared the assembly results with and without the use of reference gene annotations. We found that Cufflinks and StringTie could assemble the expected transcript consistently, without using reference gene annotations (Figure S3A). A further comparison revealed that Cufflinks had false-negative (Figure S3B) and false-positive (Figure S3C) results in assembling other representative semi-extractable RNAs, whereas StringTie yielded a stable performance. Of note, StringTie (the version used in this study) incorrectly merges two overlapping transcripts on the forward and reverse strands, respectively, into one transcript (Figure S4A). Therefore, the mapped reads were divided by the forward and reverse strands, and then, the transcripts were assembled separately before merging them into a single set (Figure S4B). For consistency and simplicity in subsequent analyses, we collapsed all transcripts assembled in each sample into a final reference transcriptome (Figure S5).

A total of 1,325 semi-extractable RNAs were identified across five human cell lines

To identify reliable semi-extractable RNAs, we eliminated transcripts with low expression levels. Transcripts with FPKM values higher than one are usually considered to be expressed in cells (54). In this study, we used a more stringent threshold (i.e., ≥ 1.5 FPKM) to screen for stably expressed transcripts. For each transcript, we quantified its semi-extractability using the expression increment of it obtained by the improved extraction method versus the conventional extraction method. A larger increment indicates higher semi-extractability of the transcript. We empirically defined transcripts with more than a 1.5-fold change in FPKM expression as semi-extractable RNAs. Finally, 187–708 semi-extractable RNAs were identified from each of the five cell lines (Figure 1B, Table S2). NEAT1 lncRNA has been reported to be the most remarkable semi-extractable RNA in HeLa cells (29). This result was reproduced using HeLa cells, as shown in (Figure 1B-C). NEAT1 was found to exhibit consistent semi-extractability in four other cell lines. Moreover, the expression level of NEAT1 was almost the highest among all semi-extractable RNAs.

We proceeded to determine whether transcripts other than NEAT1 exhibited stable semi-extractability across various cell lines. We investigated the overlap between the semi-extractable RNAs identified in A10, A549, HAP1, HEK, and HeLa cells (Figure 1D). A total of 1,526 different semi-extractable RNAs were detected in the five cell lines. Of these RNAs, most (70.45%) exhibited semi-extractability in a single cell line, reflecting somewhat cell specificity. Interestingly, we discovered that five transcripts, including NEAT1, FTX, TRIO, EXT1, and ZCCHC7, exhibited consistent and stable semi-extractability in all cell lines (Figure 1C and S6). NEAT1 and FTX are long non-coding RNAs, and the remaining three transcripts encode proteins.

Extractable RNAs were defined as RNAs with had pronounced expression changes using the improved extraction method (Figure 1B). We obtained 6,439 extractable RNAs from the five cell lines, of which 76.32% were detected as extractable in a single cell line (Figure S7 and Table S2). This result is consistent with the cell-specific levels of semi-extractable RNAs. Interestingly, Venn diagram analysis revealed 201 RNAs that exhibited semi-extractable and extractable switching between the different cell lines (Figure 1E). After removing such switching RNAs, 1,325 semi-extractable RNAs (SE) and 6,238 extractable RNAs (EX) were obtained. In addition, all unspliced transcripts

Table 1. Semi-extractable RNAs preferentially transcribe from the enhancer, repetitive, and repressed regions. The chromatin state in HeLa cells was annotated in advance using chromHMM and obtained from the ENCODE project. The percentages of various chromatin states in the transcribed regions of semi-extractable RNAs (%se), extractable RNAs (%ex), and all unspliced RNAs (%all) were calculated separately. The ratio of %se to %ex (se/ex) and %all (se/all) measures the transcriptional preference of semi-extractable RNAs in different chromatin states. Sorted by se/ex column in descending order.

ChromHMM states	%se	%ex	%all	se/ex	ex/all
Candidate Strong enhancer	2.53	1.06	1.19	2.39	2.13
Candidate Weak enhancer/DNase	5.20	2.80	4.42	1.86	1.18
Heterochromatin/Repetitive/Copy Number Variation	7.53	4.23	25.72	1.78	0.29
Polycomb repressed	0.35	0.25	5.88	1.38	0.06
Low activity proximal to active states	36.58	27.78	44.69	1.32	0.82
Promoter Flanking	1.31	1.59	0.64	0.83	2.04
Transcription associated	42.53	54.94	12.83	0.77	3.31
Distal CTCF/Candidate Insulator	0.82	1.13	1.28	0.73	0.64
Inactive Promoter	0.05	0.10	0.14	0.55	0.39
Active Promoter	2.84	5.34	1.72	0.53	1.65

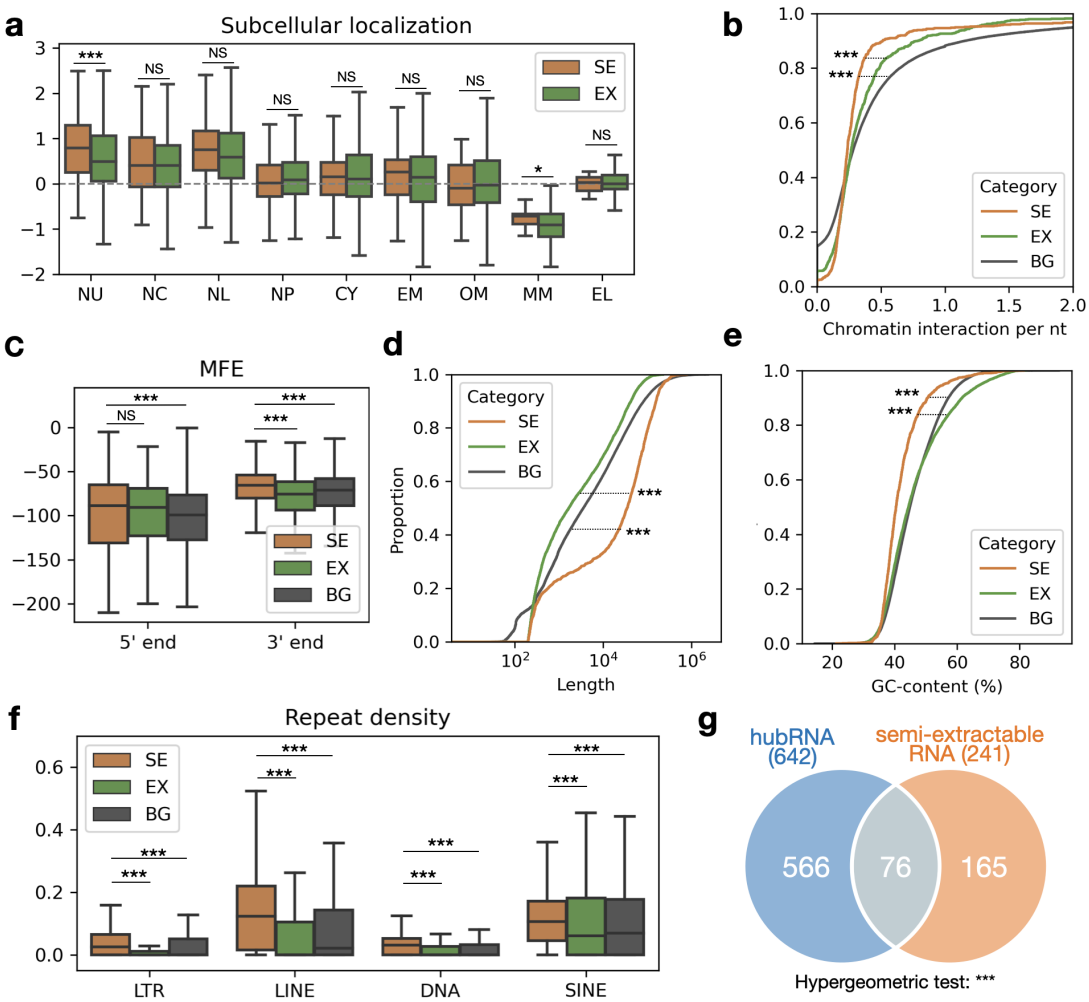

Figure 2. Characterization of semi-extractable RNAs. (a) Comparing subcellular RNA localization measured by APEX-seq fold changes in HEK cells. Increasing values indicate higher abundance in the corresponding subcellular fractions. NU: nucleolus, NC: nucleus, NL: nuclear lamina, NP: nuclear pore, CY: cytosol, EM: ER membrane, OM: outer mitochondrial membrane, MM: mitochondrial matrix, EL: ER lumen. **(b)** Comparing the minimum free energy (MFE) in the 5′ and 3′ end regions that are 300 nucleotides in length. MEF was calculated based on RNAfold. Cumulative density function analysis of **(c)** length in nucleotide, **(d)** G and C content, and **(e)** chromatin-RNA interactions measured by iMARGI in HEK cells. **(f)** Repeat elements are significantly enriched in semi-extractable RNAs. SINE: Short interspersed nuclear elements, DNA: DNA transposons, LINE: Long interspersed nuclear elements, LTR: Long terminal repeat. **(g)** Venn diagram analysis of semi-extractable RNAs and hub RNAs detected by RIC-seq in HeLa cells. ***: p-value < 0.001, **: p-value < 0.01,*: p-value < 0.05, NS: no significance (Wilcoxon rank-sum test is indicated if not otherwise specified). SE: semi-extractable RNAs, EX: extractable RNAs, BG: all background/annotated intron-containing RNAs.

were prepared from the existing gene annotations as a background/control group (BG).

Semi-extractable RNAs as a platform to provide RNA-RNA interactions

To investigate the distribution of semi-extractable RNAs in the chromatin, we compared their origins with the chromatin states downloaded from the ENCODE project (Table 1). Most of the semi-extractable (79.11%) and extractable RNAs (82.72%) were produced in transcription-associated regions and low activity domains of the genome near active elements (i.e., low activity proximal to active states). Compared to extractable RNAs, semi-extractable RNAs were more enriched in enhancers (including candidate strong/weak enhancer and weak DNase hypersensitive sites), repetitive/heterochromatin (heterochromatin/repetitive/copy number variation), and repressed (polycomb repressed and low activity proximal to active states) regions, with limited distribution in promoter regions. We argued that unspliced RNAs were inappropriate as controls, because this control group failed to account for the expressed transcripts and could not control cell specificity and epigenomic complexity. Furthermore, as the length of intron regions in unspliced RNAs is markedly larger than that of exons (55), the percentage of low activity and repetitive regions in all unspliced RNAs was more than 75%. This percentage markedly differs from that of semi-extractable and extractable RNAs. In summary, semi-extractable RNAs were preferentially derived from the functional regulatory regions of chromatin.

We proceeded to examine the subcellular localization of the semi-extractable RNAs. For each transcript, we calculated

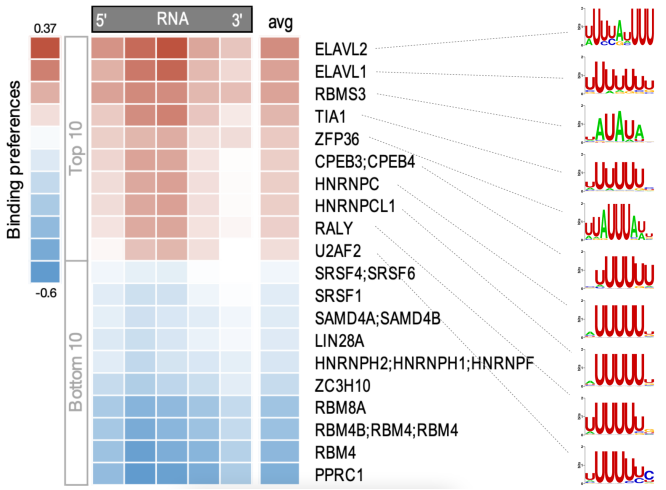

Figure 3. Motif enrichment analysis of semi-extractable RNAs. RBP binding preferences in different positional regions of semi-extractable RNAs, controlled with extractable RNAs. Here, x and y axes represent RNAs and RBPs, respectively. RNA was divided equally into five regions, and the RBP binding density within the regions was predicted using FIMO. The avg column indicates the average binding preference of RBP over the whole RNA. The result is sorted by the avg column. RBP binding sequence motifs are shown on the right. See Table S2 for details.

the degree of preference for nine subcellular fractions from publicly available APEX-seq data (42). Both semi-extractable RNAs and extractable RNAs tended to be localized in the nucleus (including the nucleolus, nucleus, and nuclear lamina) rather than in the mitochondrial matrix (Figure 2A). Interestingly, semi-extractable RNAs were significantly (p-value < 0.001) enriched in the nucleolus fraction. We further investigated the association of semi-extractable RNAs with chromatin using public iMARGI data (46). Semi-extractable RNAs were found to be disassociated from chromatin (Figure 2B). Overall, semi-extracted RNAs appear to be localized in the nucleus and are particularly enriched in the nucleolus.

NEAT1 forms paraspeckles through specific sequence features and an RNA-based interactome (26, 28). As NEAT1 was also identified as a consistent semi-extractable RNA aross cell lines in this study, we were curious whether semi-extractable RNAs possessed sequence characteristics similar to those of NEAT1. First, we determined whether the $5'$ and $3'$ ends of the semi-extractable RNAs had strong RNA structures to maintain RNA stability (28). We used the MFE of a sequence as a proxy for measuring the strength of the RNA structure. For an RNA sequence, a lower MFE indicates a higher propensity for strongly structured RNA. Surprisingly, the $5'$ and $3'$ ends of semi-extractable RNAs tended to have weak RNA structures (Figure 2C). In addition, we observed that the semi-extractable RNAs were significantly longer (Figure 2D) with lower GC content (Figure 2E) than the extractable RNAs. Repeat elements (particularly LINEs and SINEs) were significantly enriched in semi-extractable RNAs (Figure 2F). Based on the above observations, we hypothesized that semi-extractable RNAs are potential platforms for interactions with other RNAs. To test this hypothesis, we obtained 642 hub RNAs, which were detected to form RNA-RNA interactions with multiple RNAs from public RIC-seq data (47). Venn diagram analysis revealed that hub RNAs were significantly enriched (31.54%, 76 of 241) in the semi-extractable RNAs (Figure 2G).

Multifunctionality of the semi-extractable RNAs as reflected in clustered RBPs

We next explored the RBPs that bound to the semi-extractable RNAs. We downloaded the binding sequence motifs of 400 RBPs obtained by experimental validation from the CISBP-RNA database and used them to predict the binding preference of RBPs on semi-extractable RNAs. We found that RBPs that recognize AU-rich sequences were preferentially associated with semi-extractable RNAs (Figure 3 and Table S3). AU-rich elements have been reported in the $3'$ UTRs of many mRNAs and are associated with the regulation of RNA stability (56, 57). Interestingly, RBPs recognizing AU-rich elements were concentrated in the middle of the semi-extractable RNAs (Figure 3), implying that AU-rich elements in semi-extractable RNAs may be involved in other uncovered functions.

In addition, the reported paraspeckle RBPs enriched in NEAT1 (28) did not have a global binding preference for semi-extractable RNAs (Table S3). Hence, we hypothesized that the semi-extractable RNAs might contain functionally diverse RNAs, a group that possesses functions similar to that of the NEAT1 constituent paraspeckles. Accordingly, we divided the semi-extractable RNAs into four groups/communities with potentially different functions based on sequence similarity (Figure 4A). Among the five stable semi-extractable RNAs, ZCCHC7 belonged to group 1, NEAT1 and TRIO belonged to group 2, and EXT1 and FTX belonged to group 3 (Figure 4B). Furthermore, we examined the above four groups of semi-extractable RNAs for RBP-binding preference. Paraspeckle RBPs, such as FUS and NONO, preferentially bound to group 2 containing NEAT1 (Figure 4C). Theses results are consistent with those of a previous study (28).

DISCUSSION

In this study, 1,325 semi-extractable RNAs were systematically identified from five human cell lines, thereby providing an essential resource for studying RNA-centric phase separation. Biomolecular condensates without membranes are typically formed via phase separation in cells. Most previous studies have focused on the role of various proteins in forming phase-separated structures, and many proteins associated with phase separation have been explored (1, 13, 14, 15). However, numerous researchers have recently turned their attention to the role of RNA in phase separation (16, 17, 18, 19, 20, 21). NEAT1 has been reported to act as an architectural RNA to form a membrane-less condensate in the nucleus, called the paraspeckle (22, 23, 24, 25). Previous studies have experimentally verified that semi-extractable RNAs, including NEAT1, can induce the formation of nuclear bodies (29). Therefore, the RNAs contained in condensates could be poorly harvested by conventional RNA extraction and exhibited semi-extractability. Thus, the semi-extractable RNAs detected in this study may have been derived from various phase-separated condensates. Such semi-extractable RNAs may be segregated into condensates

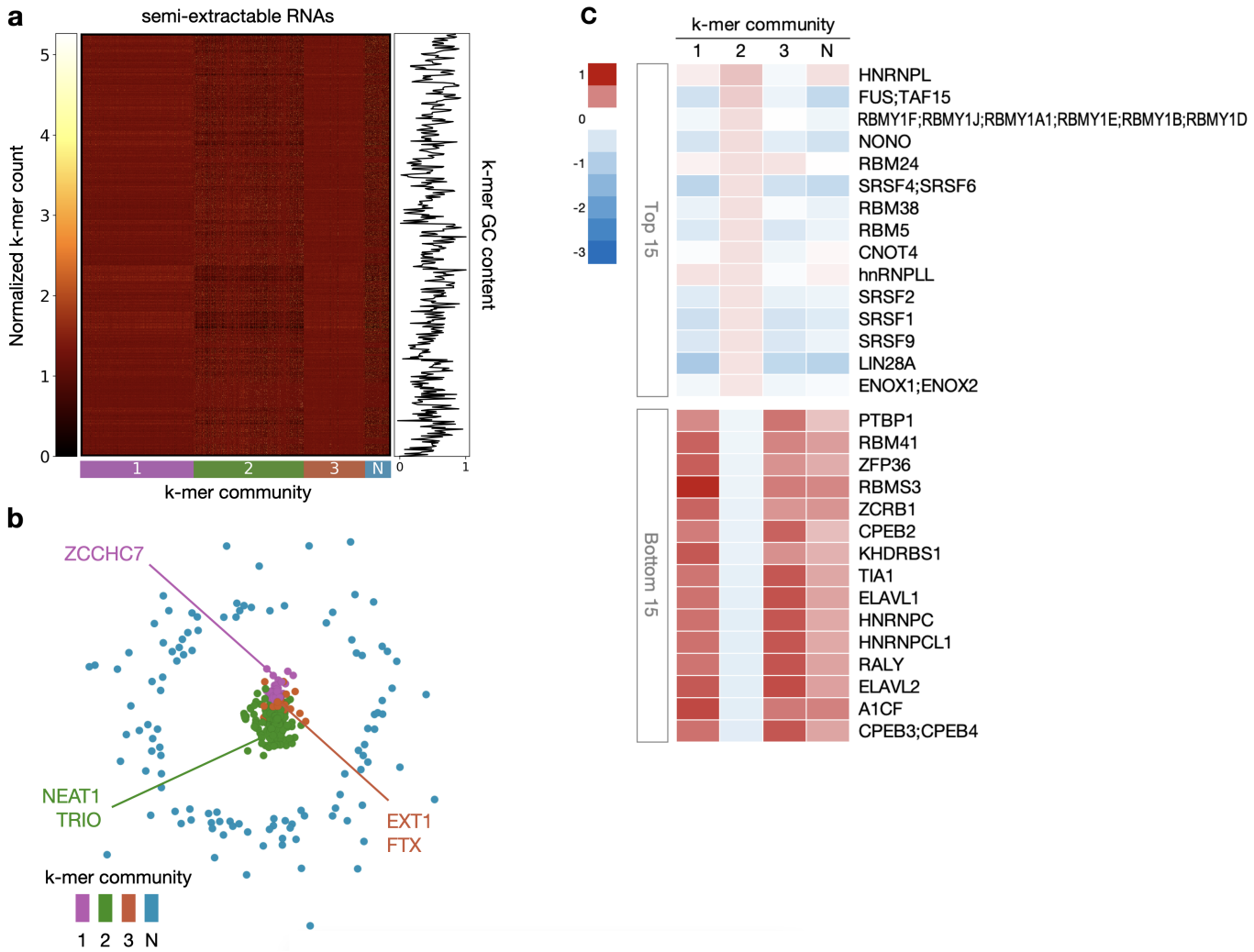

Figure 4. Clustering analysis of semi-extractable RNAs. (a) Louvain-assigned community of semi-extractable RNAs at k-mer length 6, with semi-extractable RNAs and k-mers on the x and y axes, respectively. Normalized k-mer count ranges from black (lowest) to yellow (highest). GC content of the k-mers is shown on the right panel. A side bar of the k-mer community is shown below the x axis. N means the null community. **(b)** Network graph of semi-extractable RNAs. RNA names are colored by their Louvain community assignment. **(c)** RBP-binding preference analysis was performed for each semi-extractable RNA community separately. Ordered based on the binding preference of community 2. See Table S3 for details.

by specific biological functions. However, GO analysis showed that semi-extractable RNAs were involved in a broad range of biological processes (Figure S8). We proposed the following two hypotheses to explain this result: First, the semi-extractable RNAs may be a mixture of RNAs derived from condensates with different biological functions. As semi-extractable RNAs can be further classified according to the type of condensates, the specific biological functions involving these RNAs could be identified. Second, semi-extractable RNAs may be involved in specific biological regulatory processes as RNA molecules, and these functions are not detectable by GO analysis based on protein function and phenotype annotation.

According to subcellular localization analysis of semi-extractable RNAs, semi-extractable RNAs were enriched in the nucleolus. This phenomenon is inconsistent with the previous observation that semi-extractable RNAs are primarily derived from the nuclear bodies (29); this may be due to the dynamic exchange of contents, including RNAs, between the nucleolus and nuclear bodies (3). The semi-extractable RNAs were divided into four groups that may perform different biological functions based on sequence similarity (Figure 4B). Among them, the semi-extractable RNAs in group 2, where NEAT1 is located, preferentially bind to some known paraspeckle RBPs (i.e., NONO, FUS) (Figure 4C), implying that this group of semi-extractable RNAs may possess similar functions to NEAT1 in constituting the granule backbone. PVT1, as a stable semi-extractable RNA in this group (Table S2), was observed to form a complex network as hub RNAs with other RNAs through RBP-mediated RNA-RNA interactions, which can form granule-like foci in the nucleus, and PVT1 foci do not intersect with known nuclear bodies (29, 47). PVT1 is a neighbor of the well-known oncogene, MYC, and has been reported to be involved in the regulation of cancer development (58, 59, 60, 61, 62, 63). The semi-extractable property of PVT1 implies a new aspect of phase separation for investigating its molecular regulation in the mechanism of tumorigenesis. Notably, we

detected a stable semi-extracted RNA from four cell lines, Merged.plus.5972.1 (chr17 43323708-43361338, Table S2), originating from the intergenic region. Merged.plus.5972.1 and NEAT1 were found to belong to the same group by sequence similarity. Interestingly, Merged.plus.5972.1 is located in the upstream region of the U2 small nuclear RNA (snRNA) gene cluster, suggesting that Merged.plus.5921.1 may form RNA foci and participate in U2 RNA processing. FTX in group 3 is involved in X chromosome inactivation as a positive regulator of XIST (64). This function has been reported to depend on FTX transcription, rather than its RNA product (65). However, the semi-extractability of FTX suggests that its RNA product may be involved in X chromosome inactivation via intracellular condensates. XIST has been reported to form a phase-separated compartment by interacting with multiple RBPs (20, 66, 67). However, in this study, the XIST was not observed to be semi-extractable. ZCCHC7 in group 1 is involved in RNA quality regulation after translation into proteins (68), especially viral RNA degradation (69). The semi-extractability of ZCCHC7 implies that its RNA product may be harbored in the cell as biomolecular condensates, without being eagerly used for protein production, but can rapidly respond to the invasion of pathogenic RNAs.

Numerous repetitive sequences were identified in the semi-extractable RNAs (Figure 2F, S6), which is not consistent with our speculation, as we discarded the multi-mapping reads that may result from repetitive sequences. There two potential reasons for these results. First, many reads may be mapped to nonrepetitive regions for repeat-containing RNAs, allowing the expression levels of these RNAs to be detected. Second, reads containing repetitive sequences may still be uniquely mapped owing to mutations or unique flanking sequences in the repeats. Consistently, many RNAs that contain repeats have been reported to be associated with phase separation. For example, CTN-RNA was found to be distributed in mouse paraspeckles. CTN-RNA contains three inverted repeats from SINE, which are thought to affect A-to-I editing and nuclear retention (70). CAG-repeat-containing RNA was observed to colocalize with nuclear speckles that sequester splicing factors under *in vitro* conditions (71). HSATIII lncRNAs mainly consist of primate-specific satellite III repeats, which form nuclear stress bodies under thermal stress conditions (72) and recruit specific proteins, such as heat shock factor, chromatin-remodeling complex, and splicing factors (21, 73, 74). The middle domain of NEAT1 contains repetitive sequences from LINE and SINE and this region recruits NONO dimers to trigger paraspeckle assembly (28). A systematic analysis of the potential role of repetitive sequences in the formation of RNA condensates could further our understanding of the biological mechanism of phase separation (75, 76).

Finally, we opted to discuss possible limitations and directions for further work in this study. First, we did not considering spliced isoforms when assembling the reference transcriptome, aligning with a previous report that semi-extractable RNAs are unspliced (29). However, this does not exclude the presence of spliced transcripts possessing semi-extractability under certain conditions (e.g., cell types and stress conditions). A possible solution is to combine all annotated spliced isoforms into the reference transcriptome, which would increase transcriptome complexity and affect subsequent analyses (77). Second, due to the various biases of short RNA-seq (e.g., RNA fragmentation, PCR amplification, and sequence context), the transcriptome may not be assembled accurately. We may consider adding nanopore direct RNA-seq data to assist in obtaining full-length reference transcripts (78). Third, we compared semi-extractable RNAs with public experimental data (e.g., iMARGI, RIC-seq, and APEX-seq). Of note, these experimental data involved conventional RNA extraction and thus may have lost the information on semi-extractable RNAs. Repeating the above experiments while applying improved RNA extraction is a necessary direction of work to be completed. Fourth, various stress conditions can induce the formation of different phase-separated condensates (79, 80, 81). Therefore, exploring RNA semi-extractability under various stress conditions is expected to provide important clues for our study of the potential function of phase separation in the cellular stress response (Figure S9 and Table S2). Subsequent efforts will focus on RNAs that exhibit semi-extractability under specific stimulus conditions. Finally, there is growing evidence that RNA post-transcriptional modifications can regulate the dynamics of phase separation (82, 83, 84). An interesting direction of research is to investigate whether RNA modifications are associated with the semi-extractability of RNAs.

CONCLUSION

To the best of our knowledge, this study provides the first dataset of genome-wide semi-extractable RNAs across cell lines (Table S2). This resource is expected to guide the exploration of RNA-based phase separations. Future use of semi-extractable RNAs in conjunction with RNA-centric interactome (46, 47, 85, 86, 87) will shed light on the molecular basis of the RNA-induced phase separation within cells.

ACKNOWLEDGEMENTS

Computations were partially performed on the NIG supercomputer at ROIS National Institute of Genetics.

FUNDING

This work was supported by JST CREST [grant no. JPMJCR20E6], AMED [grant no. 21479280], and JSPS KAKENHI [grants nos. 20H00448, 21H05276, and 22K19293] to TH; AMED [grant no. 21479280], JSPS KAKENHI [grant nos. 22H04925, 20H00624, 17K20032] to MH; JSPS KAKENHI [grants nos. 20K15784, 22K15093] to CZ.

Conflict of interest statement. None declared.

REFERENCES

1. Salman F Banani, Hyun O Lee, Anthony A Hyman, and Michael K Rosen. Biomolecular condensates: organizers of cellular biochemistry. *Nature Reviews Molecular Cell Biology*, 18(5):285–298, 2017.
2. Gregory L Dignon, Robert B Best, and Jeetain Mittal. Biomolecular phase separation: From molecular driving forces to macroscopic properties. *Annual Review of Physical Chemistry*, 71:53, 2020.

3. Diana M Mitrea and Richard W Kriwacki. Phase separation in biology; functional organization of a higher order. *Cell Communication and Signaling*, 14(1):1–20, 2016.
4. Andrew S Lyon, William B Peeples, and Michael K Rosen. A framework for understanding the functions of biomolecular condensates across scales. *Nature Reviews Molecular Cell Biology*, 22(3):215–235, 2021.
5. Clifford P Brangwynne, Christian R Eckmann, David S Courson, Agata Rybarska, Carsten Hoege, Jöbin Gharakhani, Frank Jülicher, and Anthony A Hyman. Germline P granules are liquid droplets that localize by controlled dissolution/condensation. *Science*, 324(5935):1729–1732, 2009.
6. Mario Cioce and Angus I Lamond. Cajal bodies: a long history of discovery. *Annual Review of Cell and Developmental Biology*, 21:105, 2005.
7. Archa H Fox and Angus I Lamond. Paraspeckles. *Cold Spring Harbor Perspectives in Biology*, 2(7):a000687, 2010.
8. Angus I Lamond and David L Spector. Nuclear speckles: a model for nuclear organelles. *Nature Reviews Molecular Cell Biology*, 4(8):605–612, 2003.
9. David SW Protter and Roy Parker. Principles and properties of stress granules. *Trends in Cell Biology*, 26(9):668–679, 2016.
10. Bin Wang, Lei Zhang, Tong Dai, Ziran Qin, Huasong Lu, Long Zhang, and Fangfang Zhou. Liquid–liquid phase separation in human health and diseases. *Signal Transduction and Targeted Therapy*, 6(1):1–16, 2021.
11. Sohum Mehta and Jin Zhang. Liquid–liquid phase separation drives cellular function and dysfunction in cancer. *Nature Reviews Cancer*, 22(4):239–252, 2022.
12. Xuhui Tong, Rong Tang, Jin Xu, Wei Wang, Yingjun Zhao, Xianjun Yu, and Si Shi. Liquid–liquid phase separation in tumor biology. *Signal Transduction and Targeted Therapy*, 7(1):1–22, 2022.
13. Vladimir N Uversky. Intrinsically disordered proteins in overcrowded milieu: Membrane-less organelles, phase separation, and intrinsic disorder. *Current Opinion in Structural Biology*, 44:18–30, 2017.
14. Anastasia C Murthy, Gregory L Dignon, Yelena Kan, Gül H Zerze, Sapun H Parekh, Jeetain Mittal, and Nicolas L Fawzi. Molecular interactions underlying liquid- liquid phase separation of the FUS low-complexity domain. *Nature Structural & Molecular Biology*, 26(7):637–648, 2019.
15. Qian Li, Xi Wang, Zhihui Dou, Weishan Yang, Beifang Huang, Jizhong Lou, and Zhuqing Zhang. Protein databases related to liquid–liquid phase separation. *International Journal of Molecular Sciences*, 21(18):6796, 2020.
16. Man Wu, Guang Xu, Chong Han, Peng-Fei Luan, Yu-Hang Xing, Fang Nan, Liang-Zhong Yang, Youkui Huang, Zheng-Hu Yang, Lin Shan, et al. lncRNA SLERT controls phase separation of FC/DFCs to facilitate Pol I transcription. *Science*, 373(6554):547–555, 2021.
17. Rena Onoguchi-Mizutani, Yoshitaka Kirikae, Yoko Ogura, Tony Gutschner, Sven Diederichs, and Nobuyoshi Akimitsu. Identification of a heat-inducible novel nuclear body containing the long noncoding RNA MALAT1. *Journal of Cell Science*, 134(10):jcs253559, 2021.
18. Marina Garcia-Jove Navarro, Shunnichi Kashida, Racha Chouaib, Sylvie Souquere, Gérard Pierron, Dominique Weil, and Zoher Gueroui. Rna is a critical element for the sizing and the composition of phase-separated RNA–protein condensates. *Nature Communications*, 10(1):1–13, 2019.
19. Tetsuro Hirose, Tomohiro Yamazaki, and Shinichi Nakagawa. Molecular anatomy of the architectural NEAT1 noncoding RNA: The domains, interactors, and biogenesis pathway required to build phase-separated nuclear paraspeckles. *Wiley Interdisciplinary Reviews: RNA*, 10(6):e1545, 2019.
20. Amy Pandya-Jones, Yolanda Markaki, Jacques Serizay, Tsotne Chitiashvili, Walter R Mancia Leon, Andrey Damianov, Constantinos Chronis, Bernadett Papp, Chun-Kan Chen, Robin McKee, Xiao-Jun Wang, Anthony Chau, Shan Sabri, Heinrich Leonhardt, Sika Zheng, Mitchell Guttman, Douglas L Black, and Kathrin Plath. A protein assembly mediates Xist localization and gene silencing. *Nature*, 587(7832):145–151, 2020.
21. Kensuke Ninomiya, Shungo Adachi, Tohru Natsume, Junichi Iwakiri, Goro Terai, Kiyoshi Asai, and Tetsuro Hirose. LncRNA-dependent nuclear stress bodies promote intron retention through SR protein phosphorylation. *The EMBO Journal*, 39(3):e102729, 2020.
22. Takeshi Chujo, Tomohiro Yamazaki, and Tetsuro Hirose. Architectural RNAs (arcRNAs): A class of long noncoding rnas that function as the scaffold of nuclear bodies. *Biochimica et Biophysica Acta (BBA)-Gene Regulatory Mechanisms*, 1859(1):139–146, 2016.
23. Yasnory TF Sasaki, Takashi Ideue, Miho Sano, Toutai Mituyama, and Tetsuro Hirose. Men ε/β noncoding RNAs are essential for structural integrity of nuclear paraspeckles. *Proceedings of the National Academy of Sciences*, 106(8):2525–2530, 2009.
24. Christine M Clemson, John N Hutchinson, Sergio A Sara, Alexander W Ensminger, Archa H Fox, Andrew Chess, and Jeanne B Lawrence. An architectural role for a nuclear noncoding RNA: NEAT1 RNA is essential for the structure of paraspeckles. *Molecular Cell*, 33(6):717–726, 2009.
25. Hongjae Sunwoo, Marcel E Dinger, Jeremy E Wilusz, Paulo P Amaral, John S Mattick, and David L Spector. Men ε/β nuclear-retained non-coding RNAs are up-regulated upon muscle differentiation and are essential components of paraspeckles. *Genome Research*, 19(3):347–359, 2009.
26. Audrey Jacq, Denis Becquet, Séverine Guillen, Bénédicte Boyer, Maria-Montserrat Bello-Goutierrez, Jean-Louis Franc, and Anne-Marie François-Bellan. Direct RNA–RNA interaction between Neat1 and RNA targets, as a mechanism for RNAs paraspeckle retention. *RNA Biology*, 18(11):2016–2027, 2021.
27. Jason A West, Mari Mito, Satoshi Kurosaka, Toru Takumi, Chiharu Tanegashima, Takeshi Chujo, Kaori Yanaka, Robert E Kingston, Tetsuro Hirose, Charles Bond, Archa Fox, and Shinichi Nakagawa. Structural, super-resolution microscopy analysis of paraspeckle nuclear body organization. *Journal of Cell Biology*, 214(7):817–830, 2016.
28. Tomohiro Yamazaki, Sylvie Souquere, Takeshi Chujo, Simon Kobelke, Yee Seng Chong, Archa H Fox, Charles S Bond, Shinichi Nakagawa, Gerard Pierron, and Tetsuro Hirose. Functional domains of NEAT1 architectural lncRNA induce paraspeckle assembly through phase separation. *Molecular Cell*, 70(6):1038–1053, 2018.
29. Takeshi Chujo, Tomohiro Yamazaki, Tetsuya Kawaguchi, Satoshi Kurosaka, Toru Takumi, Shinichi Nakagawa, and Tetsuro Hirose. Unusual semi-extractability as a hallmark of nuclear body-associated architectural noncoding rna s. *The EMBO Journal*, 36(10):1447–1462, 2017.
30. Marcel Martin. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet. journal*, 17(1):10–12, 2011.
31. Alexander Dobin, Carrie A Davis, Felix Schlesinger, Jorg Drenkow, Chris Zaleski, Sonali Jha, Philippe Batut, Mark Chaisson, and Thomas R Gingeras. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, 29(1):15–21, 2013.
32. Adam Frankish, Mark Diekhans, Irwin Jungreis, Julien Lagarde, Jane E Loveland, Jonathan M Mudge, Cristina Sisu, James C Wright, Joel Armstrong, If Barnes, Andrew Berry, Alexandra Bignell, Carles Boix, Silvia Carbonell Sala, Fiona Cunningham, Tomás Di Domenico, Sarah Donaldson, Ian T Fiddes, Carlos García Girón, Jose Manuel Gonzalez, Tiago Grego, Matthew Hardy, Thibaut Hourlier, Kevin L Howe, Toby Hunt, Osagie G Izuogu, Rory Johnson, Fergal J Martin, Laura Martínez, Shamika Mohanan, Paul Muir, Fabio C P Navarro, Anne Parker, Baikang Pei, Fernando Pozo, Ferriol Calvet Riera, Magali Ruffier, Bianca M Schmitt, Eloise Stapleton, Marie-Marthe Suner, Irina Sycheva, Barbara Uszczynska-Ratajczak, Maxim Y Wolf, Jinuri Xu, Yucheng T Yang, Andrew Yates, Daniel Zerbino, Yan Zhang, Jyoti S Choudhary, Mark Gerstein, Roderic Guigó, Tim J P Hubbard, Manolis Kellis, Benedict Paten, Michael L Tress, and Paul Flicek. Gencode 2021. *Nucleic Acids Research*, 49(D1):D916–D923, 2021.
33. Nuala A. O'Leary, Mathew W. Wright, J. Rodney Brister, Stacy Ciufo, Diana Haddad, Rich McVeigh, Bhanu Rajput, Barbara Robbertse, Brian Smith-White, Danso Ako-Adjei, Alexander Astashyn, Azat Badretdin, Yiming Bao, Olga Blinkova, Vyacheslav Brover, Vyacheslav Chetvernin, Jinna Choi, Eric Cox, Olga Ermolaeva, Catherine M. Farrell, Tamara Goldfarb, Tripti Gupta, Daniel Haft, Eneida Hatcher, Wratko Hlavina, Vinita S. Joardar, Vamsi K. Kodali, Wenjun Li, Donna Maglott, Patrick Masterson, Kelly M. McGarvey, Michael R. Murphy, Kathleen O'Neill, Shashikant Pujar, Sanjida H. Rangwala, Daniel Rausch, Lillian D. Riddick, Conrad Schoch, Andrei Shkeda, Susan S. Storz, Hanzhen Sun, Francoise Thibaud-Nissen, Igor Tolstoy, Raymond E. Tully, Anjana R. Vatsan, Craig Wallin, David Webb, Wendy Wu, Melissa J. Landrum, Avi Kimchi, Tatiana Tatusova, Michael DiCuccio, Paul Kitts, Terence D. Murphy, and Kim D. Pruitt. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Research*, 44(D1):D733–D745, 2016.
34. Fiona Cunningham, James E Allen, Jamie Allen, Jorge Alvarez-Jarreta, M Ridwan Amode, Irina M Armean, Olanrewaju Austine-Orimoloye, Andrey G Azov, If Barnes, Ruth Bennett, Andrew Berry, Jyothish Bhai, Alexandra Bignell, Konstantinos Billis, Sanjay Boddu, Lucy Brooks, Mehrnaz Charkhchi, Carla Cummins, Luca Da Rin Fioretto, Claire

Davidson, Kamalkumar Dodiya, Sarah Donaldson, Bilal El Houdaigui, Tamara El Naboulsi, Reham Fatima, Carlos Garcia Giron, Thiago Genez, Jose Gonzalez Martinez, Cristina Guijarro-Clarke, Arthur Gymer, Matthew Hardy, Zoe Hollis, Thibaut Hourlier, Toby Hunt, Thomas Juettemann, Vinay Kaikala, Mike Kay, Ilias Lavidas, Tuan Le, Diana Lemos, José Carlos Marugán, Shamika Mohanan, Aleena Mushtaq, Marc Naven, Denye N Ogeh, Anne Parker, Andrew Parton, Malcolm Perry, Ivana Piližota, Irina Prosovetskaia, Manoj Pandian Sakthivel, Ahamed Imran Abdul Salam, Bianca M Schmitt, Helen Schuilenburg, Dan Sheppard, José G Pérez-Silva, William Stark, Emily Steed, Kyösti Sutinen, Ranjit Sukumaran, Dulika Sumathipala, Marie-Marthe Suner, Michal Szpak, Anja Thormann, Francesca Floriana Tricomi, David Urbina-Gómez, Andres Veidenberg, Thomas A Walsh, Brandon Walts, Natalie Willhoft, Andrea Winterbottom, Elizabeth Wass, Marc Chakiachvili, Bethany Flint, Adam Frankish, Stefano Giorgetti, Leanne Haggerty, Sarah E Hunt, Garth R IIsley, Jane E Loveland, Fergal J Martin, Benjamin Moore, Jonathan M Mudge, Matthieu Muffato, Emily Perry, Magali Ruffier, John Tate, David Thybert, Stephen J Trevanion, Sarah Dyer, Peter W Harrison, Kevin L Howe, Andrew D Yates, Daniel R Zerbino, and Paul Flicek. Ensembl 2022. *Nucleic Acids Research*, 50(D1):D988–D995, 2022.

35. Sam Kovaka, Aleksey V Zimin, Geo M Pertea, Roham Razaghi, Steven L Salzberg, and Mihaela Pertea. Transcriptome assembly from long-read RNA-seq alignments with StringTie2. *Genome Biology*, 20(1):1–13, 2019.
36. W James Kent, Charles W Sugnet, Terrence S Furey, Krishna M Roskin, Tom H Pringle, Alan M Zahler, and David Haussler. The human genome browser at UCSC. *Genome Research*, 12(6):996–1006, 2002.
37. Heng Li, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, and Richard Durbin. The sequence alignment/map format and SAMtools. *Bioinformatics*, 25(16):2078–2079, 2009.
38. Fidel Ramírez, Friederike Dündar, Sarah Diehl, Björn A Grüning, and Thomas Manke. deepTools: a flexible platform for exploring deep-sequencing data. *Nucleic Acids Research*, 42(W1):W187–W191, 2014.
39. EA Feingold and L Pachter. The ENCODE (ENCyclopedia of DNA elements) project. *Science*, 306(5696):636–640, 2004.
40. Jason Ernst and Manolis Kellis. ChromHMM: automating chromatin-state discovery and characterization. *Nature Methods*, 9(3):215–216, 2012.
41. Aaron R Quinlan and Ira M Hall. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26(6):841–842, 2010.
42. Furqan M Fazal, Shuo Han, Kevin R Parker, Pornchai Kaewsapsak, Jin Xu, Alistair N Boettiger, Howard Y Chang, and Alice Y Ting. Atlas of subcellular RNA localization revealed by APEX-Seq. *Cell*, 178(2):473–490, 2019.
43. Anthony M Bolger, Marc Lohse, and Bjoern Usadel. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, 30(15):2114–2120, 2014.
44. Bo Li and Colin N Dewey. RSEM: accurate transcript quantification from rna-seq data with or without a reference genome. *BMC Bioinformatics*, 12(1):1–16, 2011.
45. Ronny Lorenz, Stephan H Bernhart, Christian Höner zu Siederdissen, Hakim Tafer, Christoph Flamm, Peter F Stadler, and Ivo L Hofacker. ViennaRNA Package 2.0. *Algorithms for Molecular Biology*, 6(1):1–14, 2011.
46. Zhangming Yan, Norman Huang, Weixin Wu, Weizhong Chen, Yiqun Jiang, Jingyao Chen, Xuerui Huang, Xingzhao Wen, Jie Xu, Qiushi Jin, Kang Zhang, Zhen Chen, Shu Chien, and Sheng Zhong. Genome-wide colocalization of RNA–DNA interactions and fusion RNA pairs. *Proceedings of the National Academy of Sciences*, 116(8):3328–3337, 2019.
47. Zhaokui Cai, Changchang Cao, Lei Ji, Rong Ye, Di Wang, Cong Xia, Sui Wang, Zongchang Du, Naijing Hu, Xiaohua Yu, Juan Chen, Lei Wang, Xianguang Yang, Shunmin He, and Yuanchao Xue. RIC-seq for global in situ profiling of RNA–RNA spatial interactions. *Nature*, 582(7812):432–437, 2020.
48. Charles E Grant, Timothy L Bailey, and William Stafford Noble. FIMO: scanning for occurrences of a given motif. *Bioinformatics*, 27(7):1017–1018, 2011.
49. Jessime M Kirk, Susan O Kim, Kaoru Inoue, Matthew J Smola, David M Lee, Megan D Schertzer, Joshua S Wooten, Allison R Baker, Daniel Sprague, David W Collins, Christopher R Horning, Shuo Wang, Qidi Chen, Kevin M Weeks, Peter J Mucha, and J Mauro Calabrese. Functional classification of long non-coding RNAs by k-mer content. *Nature Genetics*, 50(10):1474–1482, 2018.
50. Mathieu Bastian, Sebastien Heymann, and Mathieu Jacomy. Gephi: an open source software for exploring and manipulating networks. *Proceedings of the International AAAI Conference on Web and Social Media*, 3(1):361–362, 2009.
51. Uku Raudvere, Liis Kolberg, Ivan Kuzmin, Tambet Arak, Priit Adler, Hedi Peterson, and Jaak Vilo. g:Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update). *Nucleic Acids Research*, 47(W1):W191–W198, 2019.
52. Junichi Iwakiri, Kumiko Tanaka, Takeshi Chujo, Tomohiro Yamazaki, Goro Terai, Kiyoshi Asai, and Tetsuro Hirose. Remarkable improvement in detection of readthrough downstream-of-gene transcripts by semi-extractable RNA-sequencing. *bioRxiv*, XXX, 2022.
53. Sagar Chhangawala, Gabe Rudy, Christopher E Mason, and Jeffrey A Rosenfeld. The impact of read length on quantification of differentially expressed genes and splice junction detection. *Genome Biology*, 16(1):1–10, 2015.
54. Günter P Wagner, Koryu Kin, and Vincent J Lynch. A model based criterion for gene expression calls using RNA-seq data. *Theory in Biosciences*, 132(3):159–164, 2013.
55. Meena Kishore Sakharkar, Vincent TK Chow, and Pandjassarame Kangueane. Distributions of exons and introns in the human genome. *In Silico Biology*, 4(4):387–393, 2004.
56. Chyi-Ying A Chen and Ann-Bin Shyu. AU-rich elements: characterization and importance in mRNA degradation. *Trends in Biochemical Sciences*, 20(11):465–470, 1995.
57. Mireya Plass, Simon H Rasmussen, and Anders Krogh. Highly accessible AU-rich regions in 3′ untranslated regions are hotspots for binding of regulatory factors. *PLoS Computational Biology*, 13(4):e1005460, 2017.
58. Yuen-Yi Tseng, Branden S Moriarity, Wuming Gong, Ryutaro Akiyama, Ashutosh Tiwari, Hiroko Kawakami, Peter Ronning, Brian Reuland, Kacey Guenther, Thomas C Beadnell, Jaclyn Essig, George M Otto, M Gerard O'Sullivan, David A Largaespada, Kathryn L Schwertfeger, York Marahrens, Yasuhiko Kawakami, and Anindya Bagchi. PVT1 dependence in cancer with MYC copy-number increase. *Nature*, 512(7512):82–86, 2014.
59. Teresa Colombo, Lorenzo Farina, Giuseppe Macino, and Paola Paci. PVT1: a rising star among oncogenic long noncoding RNAs. *BioMed Research International*, 2015, 2015.
60. Jing Zhao, Peizhun Du, Peng Cui, Yunyun Qin, Cheng'en Hu, Jing Wu, Zhongwen Zhou, Wenhong Zhang, Lunxiu Qin, and Guangjian Huang. LncRNA PVT1 promotes angiogenesis via activating the STAT3/VEGFA axis in gastric cancer. *Oncogene*, 37(30):4094–4109, 2018.
61. Seung Woo Cho, Jin Xu, Ruping Sun, Maxwell R Mumbach, Ava C Carter, Y Grace Chen, Kathryn E Yost, Jeewon Kim, Jing He, Stephanie A Nevins, Suet-Feung Chin, Carlos Caldas, S John Liu, Max A Horlbeck, Daniel A Lim, Jonathan S Weissman, Christina Curtis, and Howard Y Chang. Promoter of lncRNA gene PVT1 is a tumor-suppressor DNA boundary element. *Cell*, 173(6):1398–1412, 2018.
62. Debora Traversa, Giorgia Simonetti, Doron Tolomeo, Grazia Visci, Gemma Macchia, Martina Ghetti, Giovanni Martinelli, Lasse S Kristensen, and Clelia Tiziana Storlazzi. Unravelling similarities and differences in the role of circular and linear PVT1 in cancer and human disease. *British Journal of Cancer*, 126(6):835–850, 2022.
63. Kevin Tabury, Mehri Monavarian, Eduardo Listik, Abigail K Shelton, Alex Seok Choi, Roel Quintens, Rebecca C Arend, Nadine Hempel, C Ryan Miller, Balázs Györrfy, et al. PVT1 is a stress-responsive lncRNA that drives ovarian cancer metastasis and chemoresistance. *Life Science Alliance*, 5(11), 2022.
64. Corinne Chureau, Sophie Chantalat, Antonio Romito, Angélique Galvani, Laurent Duret, Philip Avner, and Claire Rougeulle. Ftx is a non-coding rna which affects Xist expression and chromatin structure within the X-inactivation center region. *Human Molecular Genetics*, 20(4):705–718, 2011.
65. Giulia Furlan, Nancy Gutierrez Hernandez, Christophe Huret, Rafael Galupa, Joke Gerarda van Bemmel, Antonio Romito, Edith Heard, Céline Morey, and Claire Rougeulle. The Ftx noncoding locus controls X chromosome inactivation independently of its RNA products. *Molecular Cell*, 70(3):462–472, 2018.
66. Andrea Cerase, Alexandros Armaos, Christoph Neumayer, Philip Avner, Mitchell Guttman, and Gian Gaetano Tartaglia. Phase separation drives X-chromosome inactivation: a hypothesis. *Nature Structural & Molecular Biology*, 26(5):331–334, 2019.

67. Davide Cirillo, Mario Blanco, Alexandros Armaos, Andreas Buness, Philip Avner, Mitchell Guttman, Andrea Cerase, and Gian Gaetano Tartaglia. Quantitative predictions of protein interactions with long noncoding RNAs. *Nature Methods*, 14(1):5–6, 2017.
68. Sandra L Wolin and Lynne E Maquat. Cellular RNA surveillance in health and disease. *Science*, 366(6467):822–827, 2019.
69. Jerome M Molleston, Leah R Sabin, Ryan H Moy, Sanjay V Menghani, Keiko Rausch, Beth Gordesky-Gold, Kaycie C Hopkins, Rui Zhou, Torben Heick Jensen, Jeremy E Wilusz, and Sara Cherry. A conserved virus-induced cytoplasmic TRAMP-like complex recruits the exosome to target viral rna for degradation. *Genes & Development*, 30(14):1658–1670, 2016.
70. Kannanganattu V Prasanth, Supriya G Prasanth, Zhenyu Xuan, Stephen Hearn, Susan M Freier, C Frank Bennett, Michael Q Zhang, and David L Spector. Regulating gene expression through RNA nuclear retention. *Cell*, 123(2):249–263, 2005.
71. Ankur Jain and Ronald D Vale. RNA phase transitions in repeat expansion disorders. *Nature*, 546(7657):243–247, 2017.
72. Rut Valgardsdottir, Ilaria Chiodi, Manuela Giordano, Antonio Rossi, Silvia Bazzini, Claudia Ghigna, Silvano Riva, and Giuseppe Biamonti. Transcription of Satellite III non-coding RNAs is a general stress response in human cells. *Nucleic Aacids Research*, 36(2):423–434, 2008.
73. Caroline Jolly, Alexandra Metz, Jérôme Govin, Marc Vigneron, Bryan M Turner, Saadi Khochbin, and Claire Vourc'h. Stress-induced transcription of satellite III repeats. *The Journal of Cell Biology*, 164(1):25–33, 2004.
74. Tetsuya Kawaguchi, Akie Tanigawa, Takao Naganuma, Yasuyuki Ohkawa, Sylvie Souquere, Gerard Pierron, and Tetsuro Hirose. SWI/SNF chromatin-remodeling complexes function in noncoding RNA-dependent assembly of nuclear bodies. *Proceedings of the National Academy of Sciences*, 112(14):4304–4309, 2015.
75. Masahiro Onoguchi, Chao Zeng, Ayako Matsumaru, and Michiaki Hamada. Binding patterns of RNA-binding proteins to repeat-derived RNA sequences reveal putative functional RNA elements. *NAR Genomics and Bioinformatics*, 3(3):lqab055, 2021.
76. Chao Zeng, Atsushi Takeda, Kotaro Sekine, Naoki Osato, Tsukasa Fukunaga, and Michiaki Hamada. Bioinformatics approaches for determining the functional impact of repetitive elements on non-coding RNAs. In *piRNA: Methods and Protocols*, pages 315–340. Springer US, 2022.
77. Yu Hamaguchi, Chao Zeng, and Michiaki Hamada. Impact of human gene annotations on RNA-seq differential expression analysis. *BMC Genomics*, 22(1):1–12, 2021.
78. Daniel R Garalde, Elizabeth A Snell, Daniel Jachimowicz, Botond Sipos, Joseph H Lloyd, Mark Bruce, Nadia Pantic, Tigist Admassu, Phillip James, Anthony Warland, Michael Jordan, Jonah Ciccone, Sabrina Serra, Jemma Keenan, Samuel Martin, Luke McNeill, E Jayne Wallace, Lakmal Jayasinghe, Chris Wright, Javier Blasco, Stephen Young, Denise Brocklebank, Sissel Juul, James Clarke, Andrew J Heron, and Daniel J Turner. Highly parallel direct RNA sequencing on an array of nanopores. *Nature Methods*, 15(3):201–206, 2018.
79. Gregory G Fuller, Ting Han, Mallory A Freeberg, James J Moresco, Amirhossein Ghanbari Niaki, Nathan P Roach, John R Yates III, Sua Myong, and John K Kim. RNA promotes phase separation of glycolysis enzymes into yeast G bodies in hypoxia. *Elife*, 9, 2020.
80. Christiane Iserman, Christine Desroches Altamirano, Ceciel Jegers, Ulrike Friedrich, Taraneh Zarin, Anatol W Fritsch, Matthäus Mittasch, Antonio Domingues, Lena Hersemann, Marcus Jahnel, Doris Richter, Ulf-Peter Guenther, Matthias W Hentze, Alan M.Moses, Anthony A Hyman, Günter Kramer, Moritz Kreysing, Titus M Franzmann, and Simon Alberti. Condensation of Ded1p promotes a translational switch from housekeeping to stress protein production. *Cell*, 181(4):818–831, 2020.
81. Rena Onoguchi-Mizutani and Nobuyoshi Akimitsu. Long noncoding RNA and phase separation in cellular stress response. *The Journal of Biochemistry*, 171(3):269–276, 2022.
82. Ryan J Ries, Sara Zaccara, Pierre Klein, Anthony Olarerin-George, Sim Namkoong, Brian F Pickering, Deepak P Patil, Hojoong Kwak, Jun Hee Lee, and Samie R Jaffrey. m6A enhances the phase separation potential of mRNA. *Nature*, 571(7765):424–428, 2019.
83. Aleksej Drino and Matthias R Schaefer. RNAs, phase separation, and membrane-less organelles: Are post-transcriptional modifications modulating organelle dynamics? *BioEssays*, 40(12):1800085, 2018.
84. Hong Gil Lee, Jiwoo Kim, and Pil Joon Seo. N6-methyladenosine–modified RNA acts as a molecular glue that drives liquid–liquid phase separation in plants. *Plant Signaling & Behavior*, 17(1):2079308, 2022.
85. Michael Kertesz, Yue Wan, Elad Mazor, John L Rinn, Robert C Nutter, Howard Y Chang, and Eran Segal. Genome-wide measurement of RNA secondary structure in yeast. *Nature*, 467(7311):103–107, 2010.
86. Eric L Van Nostrand, Gabriel A Pratt, Alexander A Shishkin, Chelsea Gelboin-Burkhart, Mark Y Fang, Balaji Sundararaman, Steven M Blue, Thai B Nguyen, Christine Surka, Keri Elkins, Rebecca Stanton, Frank Rigo, Mitchell Guttman, and Gene W Yeo. Robust transcriptome-wide discovery of RNA-binding protein binding sites with enhanced CLIP (eCLIP). *Nature Methods*, 13(6):508–514, 2016.
87. Alessandro Bonetti, Federico Agostini, Ana Maria Suzuki, Kosuke Hashimoto, Giovanni Pascarella, Juliette Gimenez, Leonie Roos, Alex J Nash, Marco Ghilotti, Christopher JF Cameron, Matthew Valentine, Yulia A Medvedeva, Shuhei Noguchi, Eneritz Agirre, Kaori Kashi, Samudyata, Joachim Luginbühl, Riccardo Cazzoli, Saumya Agrawal, M Nicholas Luscombe, Mathieu Blanchette, Takeya Kasukawa, Michiel de Hoon, Erik Arner, Boris Lenhard, Charles Plessy, Gonçalo Castelo-Branco, Valerio Orlando, and Piero Carninci. RADICL-seq identifies general and cell type–specific principles of genome-wide RNA-chromatin interactions. *Nature Communications*, 11(1):1–14, 2020.