Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 June 2021 doi:10.20944/preprints202106.0145.v1

Bioinformatics Approaches for Determining the Functional Impact of Repetitive Elements on Non-coding RNAs

Chao Zeng [1,2]*, Atsushi Takeda [1], Kotaro Sekine [1], Naoki Osato [1], Tsukasa Fukunaga [3], Michiaki Hamada [1,2]*

[1] Faculty of Science and Engineering, Waseda University, 3-4-1 Okubo, Shinjuku-ku, 169-8555 Tokyo, Japan

[2] AIST-Waseda University Computational Bio Big-Data Open Innovation Laboratory (CBBD-OIL), 3-4-1 Okubo, Shinjuku-ku, 169-8555 Tokyo, Japan

[3] Waseda Institute for Advanced Study, Waseda University, 1-21-1 Nishi Waseda, Shinjuku-ku, 169-8050 Tokyo, Japan

*Correspondence: chao.zeng@aoni.waseda.jp (CZ) or mhamada@waseda.jp (MH)

Running title: Bioinformatics approaches to repetitive elements and ncRNAs

Key words: Repetitive element; Transposable element; Non-coding RNA; ncRNA; Functional element; Bioinformatics

Abstract

With a large number of annotated non-coding RNAs (ncRNAs), repetitive sequences are found to constitute functional components (termed as repetitive elements) in ncRNAs that perform specific biological functions. Bioinformatics analysis is a powerful tool for improving our understanding of the role of repetitive elements in ncRNAs. This chapter summarizes recent findings that reveal the role of repetitive elements in ncRNAs. Furthermore, relevant bioinformatics approaches are systematically reviewed, which promises to provide valuable resources for studying the functional impact of repetitive elements on ncRNAs.

1 Introduction

Non-coding RNA (ncRNA) is a general term for RNAs that are not translated into proteins but exert various functions in cells. Many ncRNAs are involved in a variety of cellular regulations, including transcription and translation, and contribute to the complexity of higher organisms. Among ncRNAs, long ncRNAs (lncRNAs) have recently attracted a great deal of attention ***(1)***. Similar to protein-coding RNAs (mRNAs), most lncRNAs are transcribed by RNA polymerase II, spliced, 5' capped, and 3' polyadenylated. In the case of humans, more lncRNA genes than mRNA genes are registered in several databases, such as GENCODE ***(2)***, MiTranscriptome ***(3)***, FANTOM CAT ***(4),*** and NONCODE ***(5)***. While some of these ncRNAs are related to diseases, such as cancers ***(6, 7)***, the functions of most lncRNAs remain unknown, and the elucidation of these functions is urgently needed.

Typical ncRNAs/lncRNAs have the following different characteristics from mRNAs: (1) they exhibit relatively low expression levels; (2) are expressed in a time- and tissue-specific manner; (3) are frequently nuclear localized; (4) are less conserved among species; and (5) exert their functions by interacting with various biomolecules (DNA, RNA, and protein). These characteristics make ncRNA analysis more difficult than mRNA analysis. Many repeat-derived (especially transposable element (TE)-derived) sequences are significantly enriched in lncRNAs compared with mRNAs ***(8–10)***. These repeat-derived sequences were long thought to be junk, but in recent years, strong evidence has indicated that TEs/repetitive elements are the functional components/domains of lncRNAs (for review, see ***(11–13)***). For example, recent studies have shown that the repetitive elements in lncRNAs are related to their expression ***(14, 15)***, subcellular localization ***(16),*** and others (see **Section 2**). In addition, the molecular mechanisms of repetitive elements suggest that they contribute to interactions with other molecules, such as the binding of transcription factors (TFs) ***(17, 18)***, RNA-binding proteins (RBPs) ***(19, 20)***, DNA ***(21),*** and other RNAs ***(22)*** (see **Section 3**).

Here, we focus on repetitive elements and lncRNA functions and specifically discuss bioinformatics approaches to elucidate their relationship. The following points should be noted from the perspective of bioinformatics: The first is the annotation of repetitive elements utilized in research. Since repeat-derived sequences of functional elements of ncRNAs are assumed to be a small part of the repetitive elements, methods to find the strongly conserved small portions, such as TEs, are necessary. Second, how to utilize high-throughput "omics" data, such as RNA-seq (expression information), ChIP-seq (TF binding), CLIP-seq (RNA-RBP interactions), MARIO (RNA-RNA interactions), GRID-seq (RNA-chromatin interactions), and DRIP-seq (RNA-DNA

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 June 2021 doi:10.20944/preprints202106.0145.v1

interactions), is important (see the latter sections for details). When analyzing repeat elements using these data from high-throughput sequencers, multiple-mapped reads to the reference genome/transcriptome should be carefully handled ***(23)***. This is because read sequences derived from repetitive elements may not be uniquely mapped to them. Third, it is important to know what bioinformatics methods/tools and databases are relevant for analyzing repetitive sequences and ncRNA functions. Here, we state the tools and data resources used for investigating the functional impact of repetitive elements on ncRNAs (see **Table 1**).

In this chapter, we describe bioinformatics approaches to elucidate the relationship between repetitive elements and lncRNA functions. The remainder of this paper is organized as follows: **Section 2** briefly summarizes the emerging roles of repetitive elements in ncRNAs. **Section 3** describes bioinformatics approaches for studying the role of repetitive elements in ncRNAs, including de novo identification of repetitive elements (**Section 3.1**), expression (**Section 3.2**), subcellular localization (**Section 3.3**), ncRNA-RNA interaction (**Section 3.4**), ncRNA-DNA interaction (**Section 3.5**), and ncRNA-protein interaction (**Section 3.6**). In **Section 4**, we conclude this review and discuss future research directions.

2 Emerging roles of repetitive elements in ncRNAs

Accumulating evidence suggests that repetitive sequences, including TEs and simple repeats, play an essential role in the biological functions of ncRNAs ***(11–13)***. A part (or all) of a repetitive sequence is a regulatory component (called repetitive element hereafter) that enables host ncRNAs to function through the following three interactions: RNA-RNA, RNA-DNA, and RNA-protein. This section reviews the current studies on repetitive elements in ncRNAs, including nuclear retention, biomolecular condensates, gene expression, and RNA processing and translation.

2.1 Nuclear retention

Repetitive elements usually facilitate the nuclear localization of host lncRNAs by providing anchoring sites for specific RBPs (**Fig. 1A**). Lubelsky and Ulisky screened dozens of lncRNAs and untranslated regions (UTRs) for sequences with nuclear retention and found a fragment SIRLOIN (short interspersed nuclear element (SINE)-derived nuclear RNA localization) ***(16)***. This fragment is derived from Alu elements and binds to hnRNPK proteins that can interact with splicing factors, thus, promoting the nuclear retention of host RNAs. A subsequent study revealed that hnRNPK triggers nuclear retention only at specific binding sites and sequence contexts in SIRLOIN. Additionally, two binding sites were provided in SIRLOIN for proteins SLTM and SNRNP7. Knockdown of these two proteins affects the nuclear localization of SIRLOIN-containing RNAs ***(24)***. Hacisuleyman *et al.* introduced the repeating RNA domain (RRD) from mouse *Firre* lncRNA into other cytoplasmic mRNAs and discovered that the RRD could yield a sufficient nuclear retention signal ***(25)***. Of note, this RRD has a significant species specificity with a prominent nuclear retention effect only in rodent lineages. Moreover, hnRNPU protein was observed to bind RRD and affect the nuclear localization of *Firre*. Consistently, LIF3 protein can bind SINEB2 in *Antisense Uchl1* and promote its nuclear retention ***(26)***. Carlevaro-Fita *et al.* systematically defined four TEs (L1PA16, L2b, MIRb, and MIRc) from lncRNAs associated with nuclear retention and found that their copy number was proportional to the degree of nuclear retention of host lncRNAs ***(10)***.

2.2 Biomolecular condensates

Numerous lncRNAs can form biomolecular condensates, such as paraspeckles or stress granules, in cells through repetitive sequences (**Fig. 1B**). *Neat1* acts as an architectural lncRNA involved in the formation of paraspeckles in cells. Yamazaki *et al.* found by deletion analysis that the central 8-16 kb region of *Neat1* enriched in long interspersed nuclear elements (LINEs) and SINEs is the critical domain for forming ordered paraspeckles ***(27)***. Moreover, they observed that NONO and SFPQ preferentially bind to this region, making the paraspeckle exhibit phase-separated properties. *HSATIII* consists mainly of satellite repeats and has been reported to be transcribed under heat shock. It can interact with proteins HSF1, SAFB, and hnRNPM to construct various nuclear stress bodies, which may be associated with chromatin organization and RNA splicing ***(28, 29)***. These results suggest that repetitive sequences may play a determining role in RNA-induced phase separation.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 June 2021 doi:10.20944/preprints202106.0145.v1

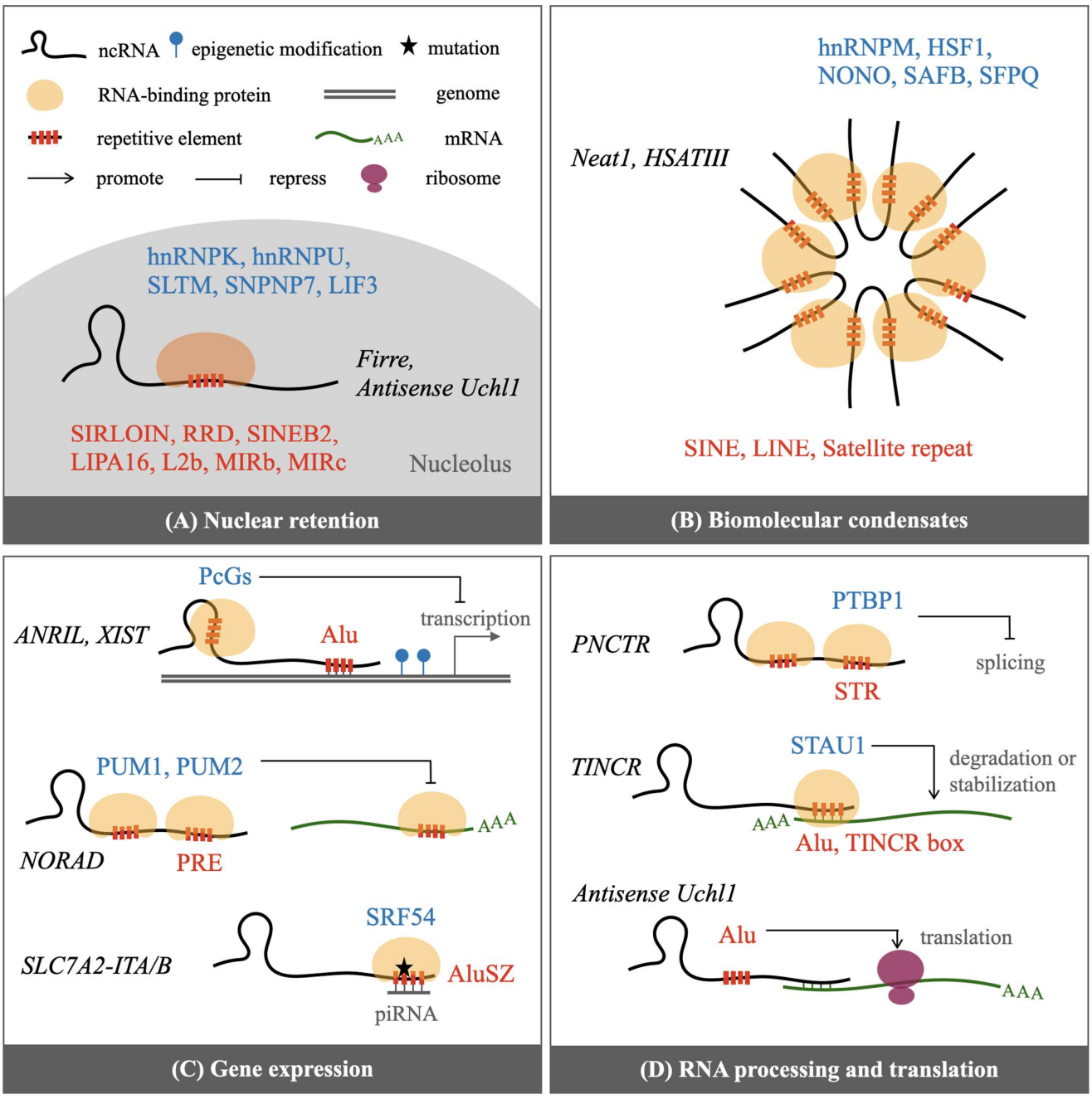

Figure 1. Repetitive elements in ncRNAs are involved in their biological functions. (A) Nuclear retention. (B) Biomolecular condensates. (C) Gene expression. (D) RNA processing and translation. Italic, blue, and red text represent ncRNAs, RNA-binding proteins, and repetitive elements, respectively.

2.3 Gene expression

LncRNAs regulate gene expression levels in a variety of ways, including chromatin association mediated by repetitive elements, epigenetic modifications, repressor interactions, and Piwi-interacting RNA (piRNA) targeting (**Fig. 1C**). Several studies have reported that Alu sequences play a vital role in the trans-acting regulation of gene expression in *ANRIL*. He et al. proposed a hypothesis model in which Alu in *ANRIL* can interact directly with target gene enhancers, allowing the Polycomb group proteins (PcGs) bound by *ANRIL* to modify the epigenetic status, thus, influencing the expression of downstream genes ***(21)***. Notably, an *in silico* study analyzed the evolution of *ANRIL* genes in 27 species and observed that TE insertions in exon 3 and exon 8 rendered

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 June 2021 doi:10.20944/preprints202106.0145.v1

high conservation of *ANRIL*, suggesting that these two TE insertions may have a biological role ***(30)***. A recent study suggested that the TE in exon 8 is a crucial factor in the genomic association that *ANRIL* possesses ***(31)***. *XIST* is one of the most widely studied lncRNAs. Wutz *et al.* revealed that the repeat A region (repA) at the 5' end of *XIST* plays a determinant role in the silencing effect of *XIST* and further showed that the two stem-loop structures formed in this repA are the key to triggering silencing ***(32)***. A subsequent study by Zhao *et al.* demonstrated that the Polycomb complex PRC2 can be recruited to repA to achieve X-chromosome inactivation ***(33)***. Tichon *et al.* found that binding sites (PREs) of Pumilio proteins (PUM1 and PUM2) were present in repeat regions of the cytoplasmic lncRNA *NORAD*. Pumillo proteins can repress the expression of mRNAs containing PREs, and *NORAD* decoys Pumillo proteins by repeats to achieve the effect of regulating the expression levels of other mRNAs ***(34)***. Cartault *et al.* observed that a point mutation in the AluSZ repeat in *SLC7A2-ITA/B* resulted in decreased expression of *SLC7A2-ITA/B* in the brain, leading to neuronal apoptosis ***(35)***. Computational analysis showed that the mutation caused the AluSZ to potentially become a piRNA target or form a stable RNA structure to recruit SRF54, a signal peptide-associated protein. Two mechanistic models elucidated the gain of the deleterious function of AluSZ after point mutation through RNA-RNA or RNA-protein interactions.

2.4 RNA processing and translation

LncRNAs are involved in repeat-induced regulation of RNA splicing, degradation, stabilization, and translation (**Fig. 1D**). *PNCTR* is a lncRNA of concern because of its increased expression in a variety of tumor cells. Yap *et al.* found that *PNCTR*, which contains multiple short tandem repeats (STRs), can aggregate multiple PTBP1 proteins in the perinuclear compartment, leading to the modulation of RNA splicing of PTBP1 to promote cell survival ***(36)***. Gong and Maquat discovered that Alu repeats in a lncRNA can form a double-stranded RNA (dsRNA) with Alu in the 3' UTR of mRNA by intermolecular base-pairing. Such dsRNAs can be recognized by STAU1 proteins that trigger mRNA degradation. Notably, since Alu repeats are widely distributed in the genome, an Alu-containing lncRNA can regulate the decay of multiple mRNAs at the same time, and an Alu-containing mRNA can also be the target of several different Alu-containing lncRNAs ***(22)***. Surprisingly, a similar molecular mechanism has been reported for *TINCR* lncRNAs to stabilize the high expression of target mRNAs. *TINCR* contains a 25 nt repetitive element called the TINCR box, which forms dsRNA with target mRNAs ***(37)***. Note that *TINCR* also functions as a protein-coding RNA and the encoded peptides affect keratinocyte keratinization ***(38)***. Carrier *et al.* reported that *Antisense Uchl1* shuttled into the cytoplasm when mTORC1 signaling was inhibited and linked to *Uchl1* by base-pairing. SINEB2 triggered the cap-independent translation of *Uchl1*, thus, promoting the translation efficiency of *Uchl1* ***(39)***. Further studies suggest that a short hairpin structure in SINEB2 may be a crucial determinant in facilitating translation ***(40)***. Interestingly, as mentioned above, SINEB2 in *Antisense Uchl1* can promote nuclear retention by binding LIF3 ***(26)***. These results suggested that repetitive elements may have distinct regulatory functions owing to their divergent subcellular localization.

3 Bioinformatics approaches for studying the role of repetitive elements in ncRNAs

3.1 *De novo* identification of repetitive elements

The identification of repetitive elements in the genome is an initial step in the repetitive element analysis. Many bioinformatics methods have been developed to discover repetitive elements, which are split into two categories: library-based and *de novo* detection. Library-based methods, including RepeatMasker ***(41)***, detect repetitive elements by searching sequence similarity against manually curated repeat sequence libraries, such as Repbase ***(42)*** and Dfam ***(43, 44)***. *De novo* detection methods, on the other hand, can find repetitive elements without the use of repeat sequence libraries. Apparently, library-based annotations have high sensitivities for known repetitive elements, whereas *de novo* detection methods can detect novel repetitive elements. The latter methods can be categorized into the following three types: tandem-repeat detection, interspersed repeat detection, and structured-repeat detection.

Tandem Repeat Finder (TRF) ***(45)*** is the most widely used program for tandem repeat detection. TRF searches for candidate regions in sequences using k-mer (substring with length of k) sliding windows, and then detects tandem repeats by aligning

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 June 2021 doi:10.20944/preprints202106.0145.v1

candidates to their surrounding sequences. Other tools, such as tantan ***(46)*** and ULTRA ***(47)***, build Hidden Markov Models (HMMs) that recognize successive repeating regions in a sequence.

There are two major strategies for interspersed repeat detection: k-mer counting and self-comparison ***(48)***. The premise behind the k-mer counting strategy is that repetitive sequences have similar k-mer profiles. The k-mer counting approach is used by several tools, including Red ***(49)***, phRAIDER ***(50),*** and P-Clouds ***(51)***. To avoid loss of sensitivities, Red utilizes HMM trained on the distribution of k-mer frequency, phRAIDER adopts spaced seeds that allow mismatches in k-mers, and P-Clouds maps k-mer clusters to the original sequence. Repeat detection based on k-mer counting is faster than that based on self-comparison. On the other hand, RECON ***(52)*** and RepeatScout ***(53)*** belonging to the self-comparison strategies use sequence alignment scores. Such methods tend to show a higher performance in the evaluation of annotated genomes. RECON, which classifies the results of exhaustive self-alignments by considering the biological characteristics of interspersed repeats, is suitable for discovering interspersed repeats with mutations ***(54)***. In several comparisons of interspersed repeat identification algorithms, RepeatScout, which finds interspersed repeats using the seed-and-extension approach proposed in BLAST ***(55)***, has demonstrated high accuracy ***(48, 56, 57)***.

Because several TE subclasses have unique structures, there are also approaches to finding structural similarities in sequences rather than sequence homology. For example, LTR_FINDER ***(58, 59)***, which is a tool for screening full-length long terminal repeat (LTR) elements, first finds repetitive regions of the terminal and then searches internal domains. Other examples are LTR_retriever ***(60)*** and LTR harvest ***(61)*** for LTRs, MITE-hunter ***(62)*** and detectMITE ***(63)*** for miniature inverted-repeat transposable elements (MITEs), and HelitronScanner ***(64)*** for Helitrons. Note that these tools will only detect the targeted subclasses and not all interspersed repeats.

The interspersed repeats were annotated using TE classification methods. TEclass ***(65)*** is a standard tool for classifying detected interspersed repeats using support vector machines (SVMs). REPCLASS ***(66)*** and PASTEC ***(67)*** combined structure-based detection and alignment with repeat sequence libraries. DeepTE ***(68)*** and TERL ***(69)*** are convolutional neural network (CNN)-based TE classification tools. There are some pipelines that perform a series of steps from interspersed repeat detection to TE annotation by combining multiple existing tools. One such pipeline is RepeatModeler2 ***(54)***, which comprises RECON, RepeatScout, LTR_harvest, and LTR_retriever. For other examples, EDTA ***(56)*** is a combination of various structure-based methods, and REPET ***(70)*** establishes TE annotations from exhaustive self-alignments (for reviews of repetitive element detection, see ***(71, 72)***).

Table 1. Resources for studying repetitive elements and ncRNAs.

Tools/data sources	URL	Description
Non-coding RNA annotations		
GENCODE ***(2)***	https://www.gencodegenes.org/	Continuously updated gene annotations for human and mouse genomes, including protein-coding and non-coding RNAs.
MiTranscriptome ***(3)***	http://mitranscriptome.org/	Gene annotations for human ncRNAs predicted from thousands of normal and cancerous RNA-seq data.
FANTOM CAT ***(4)***	https://fantom.gsc.riken.jp/cat/	Gene annotations for human lncRNAs predicted from cap analysis of gene expression (CAGE) data.
NONCODE ***(5)***	http://www.noncode.org/	Manually retrieved and integrated ncRNA database containing ncRNA gene annotations for 39 species.
LNCipedia ***(73)***	https://lncipedia.org/	Manually curated lncRNA annotations for the human genome.
Repetitive sequence annotations		
Repbase ***(42, 74)***	https://www.girinst.org/repbase/	Database of representative repetitive sequences for eukaryotic species.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 June 2021 doi:10.20944/preprints202106.0145.v1

Dfam ***(43, 44)***	https://dfam.org/	Database of transposable elements for hundreds of species.
RepeatMasker ***(41)***	https://www.repeatmasker.org/	Annotating repetitive sequences of DNA sequences.
Tandem Repeat Finder ***(45)***	https://tandem.bu.edu/trf/trf.html	Detecting tandem repeats with k-mer similarity.
tantan ***(46)***	https://gitlab.com/mcfrith/tantan	Detecting tandem repeats and low complexity sequences with HMM.
ULTRA ***(47)***	https://github.com/TravisWheelerLab/ULTRA	Detecting tandem repeats with HMM.
Red ***(49)***	http://toolsmith.ens.utulsa.edu/	Sensitive to screen transposons and simple repeats.
phRAIDER ***(50)***	https://github.com/karroje/phRAIDER	Detecting repeats based on pattern-hunter.
P-Clouds ***(51)***	http://www.evolutionarygenomics.com/ProgramsData/PClouds/PClouds.html	Detecting repeats using k-mer counts.
RECON ***(52)***	http://eddylab.org/software/recon/	Detecting and classifying repeat sequences from the genome.
RepeatScout ***(53)***	http://bix.ucsd.edu/repeatscout/	Detecting repeats by seed-extension strategy.
LTR_FINDER ***(58, 59)***	http://tlife.fudan.edu.cn/tlife/ltr_finder/	Screening LTRs using suffix-array and the Smith–Waterman algorithm.
LTR_retriever ***(60)***	https://github.com/oushujun/LTR_retriever	Highly accurate and sensitive detection of LTRs.
LTRharvest ***(61)***	http://genometools.org/tools/gt_ltrharvest.html	A flexible program for the detection of LTRs.
MITE-hunter ***(62)***	http://target.iplantcollaborative.org/mite_hunter.html	Searching miniature inverted-repeat transposable elements.
detectMITE ***(63)***	https://sourceforge.net/projects/detectmite/	MITE detection using the Lempel-Ziv complexity algorithm and CD-HIT.
HelitronScanner ***(64)***	https://sourceforge.net/projects/helitronscanner/	Detecting Helitrons with a motif-extracting algorithm.
TEclass ***(65)***	http://www.compgen.uni-muenster.de/teclass	Classifying TEs with SVM for k-mer frequencies.
REPCLASS ***(66)***	https://sourceforge.net/projects/repclass/	Classifying TEs based on sequence similarity, structural characteristics, and target site duplication.
PASTEC ***(66)***	https://urgi.versailles.inra.fr/Tools/PASTEClassifier	Classifying TEs by structural features and sequence similarities.
DeepTE ***(68)***	https://github.com/LiLabAtVT/DeepTE	Classifying TEs with convolutional neural networks (CNNs).
TERL ***(69)***	https://github.com/muriloHoracio/TERL	Classifying TEs using CNNs for transformed 2D space data.
RepeatModeler2 ***(54)***	https://github.com/Dfam-consortium/RepeatModeler	Ensemble model (LtrHarvest/Ltr_retriever, RepeatScout, and RECON) for repeat detection.
EDTA ***(56)***	https://github.com/oushujun/EDTA	Ensemble model for the generation of high-quality and non-redundant TE libraries.
REPET ***(70)***	https://urgi.versailles.inra.fr/Tools/REPET	Combined approach for the identification and classification of TEs.
Sequence alignment		
BLAST ***(55)***	https://blast.ncbi.nlm.nih.gov/Blast.cgi	The most commonly used algorithm to compare sequence similarity.
LAST ***(75)***	https://gitlab.com/mcfrith/last	Fast genome-level sequence comparison.
BWA ***(76, 77)***	http://bio-bwa.sourceforge.net/	Mapping of short and long reads against the genome.
Bowtie ***(78, 79)***	http://bowtie-bio.sourceforge.net http://bowtie-bio.sourceforge.net/bowtie2	Fast and memory-efficient mapping of short reads to the genome.
BLAT ***(80)***	http://genome.ucsc.edu/cgi-bin/hgBlat	BLAST-like alignment tool for the comparison of DNA, RNA, and proteins.
STAR ***(81)***	https://code.google.com/archive/p/rna-star/	Fast but memory-intensive mapping tool for RNA-seq data.
Tophat ***(82, 83)***	https://ccb.jhu.edu/software/tophat/	Splice-aware mapping of RNA-seq reads to the genome.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 June 2021 doi:10.20944/preprints202106.0145.v1

HISAT ***(84, 85)***	http://www.ccb.jhu.edu/software/hisat https://daehwankimlab.github.io/hisat2/	Fast and sensitive mapping of RNA-seq or DNA-seq reads to the genome.
AREM ***(86)***	https://sourceforge.net/projects/arem/	Accounting for multi-mapped reads with expectation-maximum (EM) algorithm.
RNA structure		
ViennaRNA package ***(87)***	https://www.tbi.univie.ac.at/RNA/	The most commonly used software suite for the prediction and comparison of RNA secondary structures.
CENTROIDFOLD ***(88)***	http://rtools.cbrc.jp/centroidfold/	One of the most accurate tools for the prediction of RNA secondary structures.
CapR ***(89)***	https://github.com/fukunagatsu/CapR	Calculating the structural context (stem, exterior loop, hairpin loop, bulge loop, internal loop, and multibranch loop) profiles for RNA sequences.
Rtools ***(90)***	http://rtools.cbrc.jp/	Web server for the secondary structural analysis of RNA sequences.
RNAz ***(91)***	https://www.tbi.univie.ac.at/software/RNAz/	Predicting the functional RNA secondary structures from multiple sequence alignments.
RNAforester ***(92)***	https://bibiserv.cebitec.uni-bielefeld.de/rnaforester	Calculating the RNA secondary structural similarity based on the tree alignment algorithm.
RNAmotif ***(93)***	https://github.com/dacase/rnamotif	Identifying structural motifs from RNA sequences.
IPknot ***(94)***	http://rtips.dna.bio.keio.ac.jp/ipknot/	Predicting the RNA secondary structure accounting for pseudoknots.
MXfold2 ***(95)***	https://github.com/keio-bioinformatics/mxfold2/ http://www.dna.bio.keio.ac.jp/mxfold2/	Predicting the RNA secondary structure with deep learning.
RNA-RNA, RNA-DNA, RNA-Protein interactions		
RIsearch2 ***(96)***	https://rth.dk/resources/risearch/	Predicting large-scale RNA-RNA interactions.
RIblast ***(97, 98)***	https://github.com/fukunagatsu/RIblast	Ultrafast prediction of RNA-RNA interactions based on seed-and-extension strategy.
LncRRIsearch ***(99)***	http://rtools.cbrc.jp/LncRRIsearch/	Web server for lncRNA-RNA interactome analysis.
IntaRNA ***(100)***	http://rna.informatik.uni-freiburg.de/IntaRNA/	Predicting interactions between two RNA sequences.
TargetScan ***(101)***	http://www.targetscan.org/	Predicting miRNA target sites in the genome.
piRscan ***(102)***	http://cosbi4.ee.ncku.edu.tw/pirScan/	Predicting piRNA targets for a DNA or spliced RNA sequence.
Triplexator ***(103)***	http://bioinformatics.org.au/tools/triplexator/	Searching DNA:RNA triplex structures in the genome.
TriplexFPP ***(104)***	https://github.com/yuuuuzhang/TriplexFPP	Predicting DNA:RNA triplex based on two-layer CNNs.
TDF ***(105)***	http://www.regulatory-genomics.org/tdf	Predicting RNA-DNA interactions in ncRNAs.
R-loop DB ***(106)***	http://rloop.bii.a-star.edu.sg/	Computational and experimental data of R-loop-forming sequences.
ENCODE ***(107)***	https://www.encodeproject.org/	Datasets, including eCLIP-seq of RNA-binding proteins and RNA-seq in subcellular fractions, *etc.*
MACS ***(108)***	https://github.com/macs3-project/MACS	Originally designed for peak calling for DNA-protein binding sites.
Piranha ***(109)***	https://github.com/smithlabcode/piranha	Peak calling for CLIP-seq data.
HOMER ***(110)***	http://homer.ucsd.edu/homer/ngs/peaks.html	Peak calling and motif analysis for next-gen sequencing data.
MEME ***(111)***	https://meme-suite.org/meme/	Discovering and analyzing sequence motifs.
CLAM ***(112)***	https://github.com/Xinglab/CLAM	Analyzing CLIP-seq data while accounting for multi-mapped reads.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 June 2021 doi:10.20944/preprints202106.0145.v1

3.2 Tissue/tumor-specific expression

Repetitive elements (primarily TEs) in ncRNAs are related to tissue/tumor-specific ncRNA expression. Francescatto *et al.* confirmed the enrichment of DNA/TcMar-Tigger elements in ncRNAs that are expressed specifically in brain tissue ***(113)***. Based on Nielsens' research ***(114)***, they first fitted a linear model (limma ***(115)***) to the expression data from 12 tissues, including the brain, to identify ncRNAs with tissue-specific expression. The occurrence of TEs was then compared using Fisher's exact test between ncRNAs expressed solely in brain tissue (brain-specific ncRNAs) and ncRNAs expressed in two or more different tissues (non-tissue-specific nRNAs). TE enrichment in genomes with substantial changes, such as tumor cells, has been studied via *de novo* transcriptome assembly. Attig *et al.* ***(116)*** used RNA-seq data from 31 different tumor types and *de novo* assembled transcripts using Trinity ***(117)***. They then identified ERV elements enriched in tumor-specific ncRNAs using Dfam ***(43)*** library-based TE detection. The hypergeometric test has also identified TE subclasses that are enriched in promoters of ncRNAs relative to those of mRNAs in the testis ***(118)***. According to these findings, certain TEs may be responsible for the tissue/tumor-specific expression of ncRNAs.

In tissue/tumor-specific ncRNAs, the role of repetitive elements as cis-regulatory regions has also been investigated. Previous studies have shown that certain TEs can function as tissue/tumor-specific active promoters based on annotated regions ***(8, 119, 120)***. Laurent *et al.* ***(120)*** identified very long non-coding RNAs (vlncRNAs) with tissue/tumor-specific expression based on the fold change between cell lines and primary cells. Based on the frequency of overlap between vlncRNA promoters annotated by ENCODE ***(107)*** and TE regions annotated by RepeatMasker ***(41)***, they validated the enrichment of TEs in such vlncRNA promoters.

Recently, pipelines have been developed to comprehensively analyze the relationship between tissue/tumor-specific ncRNA expression and repetitive elements. Béguec *et al.* ***(121)*** and Chishima *et al.* ***(14)*** established pipelines to comprehensively capture TEs enriched in tissue-specific ncRNAs. In these pipelines, after detecting specifically expressed lncRNAs and their tissues, the enrichment of TEs in the lncRNAs was investigated based on statistical analysis. To detect tissue-specific lncRNAs, the existing metrics were used for each. Béguec *et al.* used tau ***(122),*** which evaluates the bias of expression levels by normalizing the maximum expression levels among tissues. Chishima *et al.* adopted ROKU ***(123),*** which applies entropy and Akaike's information criterion to detect tissue-specific gene expression patterns. Furthermore, pipelines that focus on dynamic expressions have been developed. Miao *et al.* ***(124)*** and Shao and Wang ***(125)*** constructed pipelines for dynamic TE-containing ncRNA expression analysis at different developmental stages. Miao *et al*. ***(124)*** analyzed the dynamic function of TEs as transcription initiation sites using bulk ATAC-seq data at each developmental stage. Shao and Wang ***(125)*** quantified the dynamic expression of TEs at the transcript level using single-cell RNA-seq data obtained from the early stages of embryogenesis. These results revealed the dynamic regulation of TE expression in the preimplantation stage and demonstrated the tissue specificity of TE-containing ncRNA transcripts in early embryogenesis.

3.3 Subcellular localization

Subcellular localization of lncRNAs predicted by sequence features provides essential clues for analyzing and understanding the biological functions of lncRNAs. Although we have observed that some repetitive elements regulate the localization of lncRNAs ***(10, 16, 24, 25)***, studies that incorporate repetitive elements in the prediction of RNA localization are poorly undertaken.

Hamilton *et al.* proposed an analytical pipeline for identifying RNA secondary structure elements in the genome to detect localization-related sequence features ***(126)***. First, they extracted two similar RNA stem-loop structures from the well-studied GLS (*grk* localization signal) and ILS (*I* factor localization signal) in *Drosophila*. This structure is critical for recognition by the components of the Dynein-dependent localization machinery. After obtaining the sequences by sliding a window over the genome, the sequences were converted into structural data using RNALfold ***(127)***. Finally, RNAdistance ***(128)***, RNAforester ***(92),*** and RNAmotif ***(93)*** were used separately to compare the similarity between the structure data and GLS and ILS stem-loops. They found that G2 and Jockey repeats could form structures similar to GLS and ILS stem-loop structures and validated them using injection assays through which they could induce specific localization in the oocyte. The above approach can encode RNA secondary structure features in repeat elements when predicting lncRNA localization.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 June 2021 doi:10.20944/preprints202106.0145.v1

Sequencing technologies have allowed us to obtain genome-wide data for mapping lncRNAs to different subcellular compartments. Using these data, we can analyze and characterize the localization of lncRNAs in terms of their sequence features. Zeng *et al.* exploited ribosome profiling data to define over 1,000 ribosome-associated and ribosome-free lncRNAs in humans and mice ***(129)***. Then, ~100 sequence-related features containing repetitive sequence content were encoded from these lncRNAs. An L1 regularized logistic regression model ***(130)*** was used to fit these data to assess various features of ribosome association. Finally, they found that lncRNAs containing LTR repeats were more likely to bind the ribosome, whereas those lncRNAs composing LINE or SINE were more likely to be ribosome-free ***(131)***. Similarly, Nadel *et al.* investigated the importance of repetitive elements in chromatin association ***(132)***. They identified DNA:RNA hybrids and density from RNA:DNA immunoprecipitation (RDIP) data in HEK 293T cells. The L1 regression model was used to fit these data and extract crucial sequence features. Sequentially, they found that LINE could facilitate chromatin association.

3.4 ncRNA-RNA interactions

RNA-RNA interactions based on complementary base pairings are essential mechanisms of action for many ncRNAs. RNA-RNA interactions are more likely to increase target specificity than RNA-protein interactions, and repetitive sequences are important elements for forming the interaction regions between two RNAs. For example, many piRNAs, short RNAs binding with PIWI proteins, have sequences complementary to transposons. These piRNAs silence transposon activities through RNA-RNA interaction with the transposons in animal germ cells to protect the genome from destruction ***(133)***. Another example is microRNA (miRNA). miRNAs are small (approximately 20 nt) RNAs in eukaryotes that suppress the expression of target mRNAs by binding the 3' UTRs of mRNAs. Some miRNAs and miRNA target sites are derived from transposons, which regulate the expression of various genes, including housekeeping genes, in humans ***(134, 135)***. Furthermore, a transposon acts as a competing endogenous RNA (ceRNA), which binds to miRNAs and maintains mRNA expression by preventing miRNAs from binding to mRNAs. Cho and Paszkwoski discovered a transposon that works as a ceRNA of miRNA171, contributing to root development in rice ***(136)***. As reviewed in **Section 2**, some lncRNAs exert their functions by interacting with other RNAs through intrinsic repetitive elements ***(22, 37)***. Controlling mRNA expression based on RNA-RNA interactions between repetitive elements, as in these cases, seems to be a more common mechanism. Nguyen *et al.* discovered many transposon-mRNA interactions experimentally and found that the interaction regions in mRNAs were more evolutionarily conserved than the neighboring regions. These results indicated that these interactions have some biological functions ***(137)***.

Experimental or computational identification of RNA-RNA interactions is a powerful approach to discover novel repetitive element-associated RNA-RNA interactions. Recently, experimental methods for exhaustive *in vivo* RNA-RNA interaction detection based on high-throughput sequencing have been developed; for example, COMRADES ***(138)***, PARIS ***(139, 140)***, and RIC-seq ***(141)***. These methods first concatenate interaction regions by cross-linking and proximate ligation, followed by reverse transcription and sequencing the concatenated RNAs. The sequencing reads were aligned by fast RNA-seq aligners, such as STAR ***(81)***, and only gapped reads or chiastic mapping reads were extracted (*i.e.*, normally mapping reads were removed from the analysis). Finally, high-confidence interaction regions were identified by greedy assembling the remaining reads as duplex groups. RNA-RNA interactions identified by these experiments before 2017 have been registered in the RISE database, and the interaction regions can be easily searched using the web interface of the database ***(142)***.

Although these experimental methods can detect RNA-RNA interactions with high accuracy, these methods cannot identify interactions involving transcripts with tissue-specific or cell-type-specific expression patterns unless researchers perform the experiments on a particular tissue or cell type. Specifically, as many lncRNAs show tissue-specific expression patterns ***(143)***, extensive experiments are required to reveal the whole picture of lncRNA-related RNA-RNA interactions. On the other hand, computational methods can predict the interaction of any RNA, including artificial RNAs that do not exist in nature. However, the prediction accuracy is still not sufficiently high, which means that experimental and computational methods are complementary approaches. For the large-scale interaction prediction of general RNA-RNA interactions, two fast software products, RIsearch2 ***(96)*** and RIblast ***(97, 98)***, have been developed. Additionally, by using the LncRRIsearch web service, we can search predicted human

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 June 2021 doi:10.20944/preprints202106.0145.v1

and mouse lncRNA-RNA interactions by RIblast and investigate tissue-specific or subcellular localized RNA interactions ***(99)***. To detect the RNA-RNA interactions of a specific class of RNA, it is better to use specialized tools for the RNA class. This is because these tools are more accurate when using interaction rules specific to the RNA class. Some examples of such methods are TargetScan for miRNAs ***(101)*** and piRscan for *C. elegans* piRNAs ***(102)***.

3.5 ncRNA-DNA interactions

Repetitive sequences can contribute to ncRNA chromatin associations as guides or cofactors. In addition to providing direct RNA-DNA interactions, repetitive sequences can also indirectly trigger ncRNA-chromatin association by binding to RBPs. For example, a technique called silica particle-assisted chromatin enrichment (SPACE) has detected hundreds of RBPs bound to chromatin in mouse embryonic stem (mES) cells ***(144)***. To date, few studies have been conducted on repetitive sequences and ncRNA-DNA interactions that can be briefly categorized as computation-, experiment-, and hybrid-driven.

Computation-driven approaches can predict lncRNA-chromatin associations and analyze the contribution of repetitive sequences in this context ***(145)***. A lncRNA can be directly bound to a specific region of chromatin through base pairing and affects gene expression proximal to that region. Based on this assumption, Deforges *et al.* predicted candidates for trans-acting lncRNA-chromatin associations using sequence similarity in *Arabidopsis thaliana*. Considering that lncRNAs are also associated with promoters, they extracted all mRNAs with regions containing 2 kb upstream, 5' UTR, exons, introns, and 3' UTR, and then utilized BLAST ***(55)*** to retrieve lncRNAs with more than 100 nt hits in these regions. Furthermore, hundreds of lncRNA-chromatin associations were identified by positive or negative correlations in expression between lncRNAs and target mRNAs in different samples. Note that a single lncRNA can correspond to multiple distinct chromatin regions. Repetitive sequences may contribute to this multiple-mapping relationship. An intriguing example is the *XLOC_000322* lncRNA bearing SINE repeats, which is predicted to exhibit positive or negative expression correlations for 13 targets. Among these targets, *AT4G04930*, *AT3G234300*, and *AT2G03340* were validated by protoplast transformation to correlate with *XLOC_000322* in terms of expression. Remarkably, this method is not appropriate for predicting cis-acting lncRNA-chromatin associations because of the sequence complementarity between nascent lncRNAs and their loci.

Experiment-driven approaches use high-throughput techniques to detect global RNA-DNA interactions and observe whether repetitive sequences appear remarkably abundant in RNA-DNA-interacting regions employing appropriate controls. The rationale is that the enrichment of certain sequences in these regions means that these specific sequences are subject to evolutionary pressure to undergo selection due to some function. Bonetti *et al.* developed the RADICL-seq (RNA and DNA-interacting complexes ligated and sequenced) technique to probe RNA-chromatin interactions ***(146)***. In mES cells, nearly 300,000 RNA-DNA-interacting loci were detected using RADICL-seq. Interestingly, more than 95% of the RNAs involved in chromatin association containing small nuclear RNA belong to trans interactions. However, SINE, LINE, and LTR were more likely to appear in a specific pattern in RNAs mapping to cis interactions. Compared with interactions free of repetitive sequences, SINE was enriched in interactions with RNA and DNA distances ranging from 10 kb to 1 Mb, while LINE and LTR were more likely to appear in RNA-chromatin associations of long-range intervals (> 100 kb). Likewise, Zeng *et al.* extracted R-loops (structure of a DNA:RNA hybrid and a displaced DNA) from publicly available DRIP-seq (DNA-RNA immunoprecipitation followed by high-throughput DNA sequencing) data. Considering the distribution of repetitive sequences in the genome, the length and locus-specific distribution of R-loops, and the tendency of R-loops to form in nascent RNAs, the authors established separate control groups to assess the enrichment of repetitive sequences in R-loop-forming regions ***(147)***.

Hybrid-driven approaches combine multiple experimental data to validate the predicted results on a computation-driven basis. Bai *et al.* used a hybrid approach to reveal the contributing role of Alu sequences in enhancer-promoter interactions (EPIs) ***(148)***. First, BLAST ***(55)*** was used to establish whether interactions between enhancers and promoters might be formed by sequence similarity. Sequence-related associations were predicted to be between ~30% of the enhancers and ~40% of the promoters in the human genome. Then, Alu-derived sequence motifs were detected by MEME analysis ***(111)*** to be enriched in these EPIs. Intriguingly, Alu depletion was found in these EPIs by comparing the content of Alu in the genome, which was interpreted as serving

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 June 2021 doi:10.20944/preprints202106.0145.v1

as a cis-regulatory element that might be subject to some evolutionary restriction. Subsequently, the involvement of Alu repeats in the regulation of gene expression through EPIs was validated using gene and allelic expression data. To elucidate the mechanism by which (*e.g.*, DNA-DNA, RNA-DNA) Alu repeats are mediated in EPI, ChIA-PET ***(149)***, GRID-seq ***(150),*** and iMARGI ***(151)*** for RNA-DNA, Triplexator ***(103)*** prediction data for DNA:RNA triplex, and ssDRIP-seq ***(152)*** data for R-loop were analyzed. The authors concluded that Alu was involved in the construction of the EPI network by the trans-acting R-loop of promoter and enhancer RNA. Additionally, they found a co-evolutionary relationship between Alus in the enhancer and promoter, further evidencing that Alu plays a regulatory role in the formation of EPI networks.

3.6 ncRNA-Protein interactions

The human genome comprises approximately 45–60% of TEs ***(19, 153)***. TEs with high sequence similarities are associated with many genomic regions with a regulatory network. Some TE-derived sequences are inserted into RNAs (especially lncRNAs), and RBPs can recognize and bind to these sequences. RNA-RBP interactions can be mapped by CLIP-seq, which uses a strand-specific library and applies a series of databases and algorithms, such as TopHat, GENCODE, CLIP-seq peak calling, and AREM ***(82, 86, 154, 155)***. CLIP-seq reads of specific RBPs were aligned to many TE families using a combination of RepeatMasker, BEDTools, DFAM profile HMM, HMMer, and PoSSuM ***(41, 43, 156–159)***. On average, 12.2% of reads in enhanced CLIP (eCLIP) experiments included repetitive elements annotated by RepBase ***(19, 74)***. Between lncRNAs and mRNAs, RBP-TE associations were similarly enriched or deficient in exons and introns. For STAU1 binding to the Alu sequence, STAU1 sequences were enriched by 3.2–4.1-fold in Alu sequences. For hnRNA C, hnRNP C sequences were 2.3-fold enriched in antisense Alu elements in transcripts. Hundreds of enrichments of sequences were detected in RBP-TE associations. Many RBPs tend to bind to specific sequences and/or structures. The enrichment of RBP-TE pairs showed a clear tendency for RBP to bind to particular subregions within the TE. For instance, hnRNP H1 CLIP-seq reads were aligned to two specific subregions of antisense L2 elements. The conservation of RBP motifs was analyzed using PhyloP ***(160)***, and a high mutation rate across RBP motifs was detected. Many non-repetitive sequences in transcripts seemed to accumulate mutations in RBP motifs. For most RBP motifs, the coverage of CLIP-seq alignments increased in motif instances outside of the repeats. The motifs were conserved significantly in the non-repetitive 3' UTR, intron and lncRNA sequences. Therefore, TE-derived instances of the motifs potentially intercept the sequences of RBPs for the motifs. To examine whether TE binding sites have similar functions to non-repetitive sites, hnRNP C knockdown experiments were performed to define bound and unbound genes from CLIP-seq alignment coverage using peak calling strategies ***(161–163)***. The cumulative distributions of the values of the statistical tests of differential expression by Cuffdiff were plotted for genes bound only in non-repetitive sequences or only in TEs ***(164)***. The expression of genes found only in non-repetitive sequences and only in TEs was similarly increased, which was observed separately for mRNAs and lncRNAs. TE-derived and non-repetitive RBP binding sequences affect the RNA state similarly in RBP knockdown gene expression analyses.

A soft-clustering non-negative matrix factorization (NMF) method for clustering CLIP-seq peaks for RBPs was developed ***(165)***. Soft clustering clusters one RBP into multiple groups, which is necessary for RBP clustering, because many RBPs have several biological functions through binding with cofactor proteins. Conventional hierarchical clustering using cut-tree methods, such as dynamic tree, dynamic hybrid, and static, cannot cluster CLIP-seq peaks properly. For example, the NMF method identified 18 RBP groups, although the conventional methods found only five groups, which were included in the 18 groups. Many known interactions were found using only the NMF method. It also detects binding sequences such that the signal of one RBP peak is weak, and the others are strong, because it takes into account the whole binding strength of the group of RBPs.

Because sequence reads aligned to repetitive sequences were not used to remove multi-mapped reads in the conventional analyses of CLIP-seq data, the frequency of RBP binding to repeat-derived RNA sequences was underestimated ***(20)***. To identify the functional elements of repetitive sequences, subfamily- and nucleotide-based analyses are required using the eCLIP data of repeat-derived RNAs. Novel components of RBP complexes were predicted to regulate the expression of LINE1 sense strand from the analysis of eCLIP data using STAR, Piranha, RepBase, and MACS ***(74, 81, 108, 109)***. The 3' UTR of L1PA subfamilies contained putative functional elements associated with heterochromatin formation. New candidate components of the splicing complex were

found to bind to LINE1 antisense sequences. This method would be useful for predicting functional RNA elements from repeat sequences, including transposons. Previous studies have not focused on the precise patterns of RBP binding to repeat sequences, particularly TE and its subfamily. This study focused on the patterns of RBP binding to TEs with nucleotide resolution and discovered several short RNA fragments that bind to multiple RBPs and form RBP clusters. RBP binding sites and RNA secondary structures of the short RNA fragments were predicted and evaluated using Rtools, CENTROIDFOLD, CapR, ViennaRNA package, uShuffle, and RNAz ***(87–91, 166)***. These sequences can form stable stem-loop structures.

Without using long sequence similarity, a kmer-based comparison of lncRNA sequences has been proposed and developed ***(167)***. LncRNAs with the same or similar functions would have sequence similarities, even if sequence alignment algorithms do not identify similarities. The idea is based on the following features of lncRNAs: first, most lncRNAs do not have catalytic activity, and their function is affected by proteins bound to lncRNAs in cells. Second, proteins bind to RNA through 3–8 bases of motifs (k-mers). Third, for a lncRNA, the positions of motifs may not be important for its function. The existence of these motifs may be sufficient for its function, which does not require long sequence similarity.

4 Concluding Remarks

In this review, we summarize the functional roles of repetitive elements in lncRNAs, especially from a bioinformatics viewpoint. Based on bioinformatics analyses of various omics data, we have provided accumulated evidence that repetitive elements contribute to the expression, subcellular localization, binding of other molecules, and so forth. In these studies, basic computational tools, such as read mapping, peak calling, motif detection, RNA secondary structure predictions, and RNA-RNA interaction predictions, as well as databases, such as repeat annotation and lncRNA annotation, play essential roles (cf. **Table 1**). In the future, further comprehensive analyses integrating large-scale experimental data, bioinformatics tools, and databases will become more important. Bioinformatics techniques needed for further research include methods for finding shorter remnants of repetitive elements with high sensitivity and for more careful handling of multi-map reads (for a review, see ***(23)***), which are derived from repetitive sequences, and methods that enable integrative analyses of several omics data. Additionally, several repetitive elements tend to be inserted into one transcript; similar to protein domains, it is important to consider not only a single repetitive element but also the combination of elements to understand their functional relationships.

Acknowledgments

This work was supported by JSPS KAKENHI [grant numbers JP20K15784 to CZ; 16H06279, 16H05879, 17K20032, and JP20H00624 to MH].

References

1. Mercer TR, Dinger ME, and Mattick JS (2009) Long non-coding RNAs: insights into functions. Nat Rev Genet 10:155–159
2. Frankish A, Diekhans M, Ferreira A-M, et al (2019) GENCODE reference annotation for the human and mouse genomes. Nucleic Acids Res 47:D766–D773
3. Iyer MK, Niknafs YS, Malik R, et al (2015) The landscape of long noncoding RNAs in the human transcriptome. Nat Genet 47:199–208
4. Hon C-C, Ramilowski JA, Harshbarger J, et al (2017) An atlas of human long non-coding RNAs with accurate 5' ends. Nature 543:199–204
5. Zhao L, Wang J, Li Y, et al (2021) NONCODEV6: an updated database dedicated to long non-coding RNA annotation in both animals and plants. Nucleic Acids Res 49:D165–D171
6. Nguyen TM, Alchalabi S, Oluwatoyosi A, et al (2020) New twists on long noncoding RNAs: from mobile elements to motile cancer cells. RNA Biol 17:1535–1549
7. Bao Z, Yang Z, Huang Z, et al (2019) LncRNADisease 2.0: an updated database of long non-coding RNA-associated diseases. Nucleic Acids Res 47:D1034–D1037
8. Kelley D and Rinn J (2012) Transposable elements reveal a stem cell-specific class of long noncoding RNAs. Genome Biol 13:R107
9. Kapusta A, Kronenberg Z, Lynch VJ, et al (2013) Transposable elements are major contributors to the origin, diversification, and regulation of vertebrate long noncoding RNAs. PLoS Genet 9:e1003470
10. Carlevaro-Fita J, Polidori T, Das M, et al (2019) Ancient exapted transposable elements promote nuclear enrichment of human long noncoding RNAs. Genome Res 29:208–222
11. Fort V, Khelifi G, and Hussein SMI (2021) Long non-coding RNAs and transposable elements: A functional relationship. Biochim Biophys Acta Mol Cell Res 1868:118837

12. Ali A, Han K, and Liang P (2021) Role of Transposable Elements in Gene Regulation in the Human Genome. Life 11
13. Johnson R and Guigó R (2014) The RIDL hypothesis: transposable elements as functional domains of long noncoding RNAs. RNA 20:959–976
14. Chishima T, Iwakiri J, and Hamada M (2018) Identification of Transposable Elements Contributing to Tissue-Specific Expression of Long Non-Coding RNAs. Genes 9
15. Lynch VJ, Leclerc RD, May G, et al (2011) Transposon-mediated rewiring of gene regulatory networks contributed to the evolution of pregnancy in mammals. Nat Genet 43:1154–1159
16. Lubelsky Y and Ulitsky I (2018) Sequences enriched in Alu repeats drive nuclear localization of long RNAs in human cells. Nature 555:107–111
17. Sundaram V, Cheng Y, Ma Z, et al (2014) Widespread contribution of transposable elements to the innovation of gene regulatory networks. Genome Res 24:1963–1976
18. Sundaram V and Wysocka J (2020) Transposable elements as a potent source of diverse cis-regulatory sequences in mammalian genomes. Philos Trans R Soc Lond B Biol Sci 375:20190347
19. Van Nostrand EL, Pratt GA, Yee BA, et al (2020) Principles of RNA processing from analysis of enhanced CLIP maps for 150 RNA binding proteins. Genome Biol 21:90
20. Masahiro O, Chao Z, Yukiteru O, et al (2021) Binding patterns of RNA binding proteins to repeat-derived RNA sequences reveal putative functional RNA elements. NAR Genom Bioinform (in press)
21. Holdt LM, Hoffmann S, Sass K, et al (2013) Alu elements in ANRIL non-coding RNA at chromosome 9p21 modulate atherogenic cell functions through trans-regulation of gene networks. PLoS Genet 9:e1003588
22. Gong C and Maquat LE (2011) lncRNAs transactivate STAU1-mediated mRNA decay by duplexing with 3' UTRs via Alu elements. Nature 470:284–288
23. Deschamps-Francoeur G, Simoneau J, and Scott MS (2020) Handling multi-mapped reads in RNA-seq. Comput Struct Biotechnol J 18:1569–1576
24. Lubelsky Y, Zuckerman B, and Ulitsky I (2021) High-resolution mapping of function and protein binding in an RNA nuclear enrichment sequence. EMBO J e106357
25. Hacisuleyman E, Shukla CJ, Weiner CL, et al (2016) Function and evolution of local repeats in the Firre locus. Nat Commun 7:11021
26. Fasolo F, Patrucco L, Volpe M, et al (2019) The RNA-binding protein ILF3 binds to transposable element sequences in SINEUP lncRNAs. FASEB J 33:13572–13589
27. Yamazaki T, Souquere S, Chujo T, et al (2018) Functional Domains of NEAT1 Architectural lncRNA Induce Paraspeckle Assembly through Phase Separation. Mol Cell 70:1038–1053.e7
28. Jolly C, Metz A, Govin J, et al (2004) Stress-induced transcription of satellite III repeats. J Cell Biol 164:25–33
29. Aly MK, Ninomiya K, Adachi S, et al (2019) Two distinct nuclear stress bodies containing different sets of RNA-binding proteins are formed with HSATIII architectural noncoding RNAs upon thermal stress exposure. Biochem Biophys Res Commun 516:419–423
30. He S, Gu W, Li Y, et al (2013) ANRIL/CDKN2B-AS shows two-stage clade-specific evolution and becomes conserved after transposon insertions in simians. BMC Evol Biol 13:247
31. Alfeghaly C, Sanchez A, Rouget R, et al (2021) Implication of repeat insertion domains in the trans-activity of the long non-coding RNA ANRIL. Nucleic Acids Res 49:4954–4970
32. Wutz A, Rasmussen TP, and Jaenisch R (2002) Chromosomal silencing and localization are mediated by different domains of Xist RNA. Nat Genet 30:167–174
33. Zhao J, Sun BK, Erwin JA, et al (2008) Polycomb proteins targeted by a short repeat RNA to the mouse X chromosome. Science 322:750–756
34. Tichon A, Gil N, Lubelsky Y, et al (2016) A conserved abundant cytoplasmic long noncoding RNA modulates repression by Pumilio proteins in human cells. Nat Commun 7:12209
35. Cartault F, Munier P, Benko E, et al (2012) Mutation in a primate-conserved retrotransposon reveals a noncoding RNA as a mediator of infantile encephalopathy. Proc Natl Acad Sci U S A 109:4980–4985
36. Yap K, Mukhina S, Zhang G, et al (2018) A Short Tandem Repeat-Enriched RNA Assembles a Nuclear Compartment to Control Alternative Splicing and Promote Cell Survival. Mol Cell 72:525–540.e13
37. Kretz M, Siprashvili Z, Chu C, et al (2013) Control of somatic tissue differentiation by the long non-coding RNA TINCR. Nature 493:231–235
38. Eckhart L, Lachner J, Tschachler E, et al (2020) TINCR is not a non-coding RNA but encodes a protein component of cornified epidermal keratinocytes. Exp Dermatol 29:376–379
39. Carrieri C, Cimatti L, Biagioli M, et al (2012) Long non-coding antisense RNA controls Uchl1 translation through an embedded SINEB2 repeat. Nature 491:454–457
40. Podbevšek P, Fasolo F, Bon C, et al (2018) Structural determinants of the SINE B2 element embedded in the long non-coding RNA activator of translation AS Uchl1. Sci Rep 8:3189
41. Smit, AFA, Hubley, R & Green, P. RepeatMasker Open-4.0, http://www.repeatmasker.org. Accessed 1 May 2021
42. Jurka J, Kapitonov VV, Pavlicek A, et al (2005) Repbase Update, a database of eukaryotic repetitive elements. Cytogenet Genome Res 110:462–467
43. Wheeler TJ, Clements J, Eddy SR, et al (2013) Dfam: a database of repetitive DNA based on profile hidden Markov models. Nucleic Acids Res 41:D70–82
44. Storer J, Hubley R, Rosen J, et al (2021) The Dfam community resource of transposable element families, sequence models, and genome annotations. Mob DNA 12:2
45. Benson G (1999) Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res 27:573–580
46. Frith MC (2011) A new repeat-masking method enables specific detection of homologous sequences. Nucleic Acids Res 39:e23
47. Olson D and Wheeler T (2018) ULTRA: A Model Based Tool to Detect Tandem Repeats. ACM BCB 2018:37–46
48. Rodriguez M and Makałowski W (2021), Software Evaluation for de novo Detection of Transposons, bioRxiv http://dx.doi.org/10.1101/2021.02.08.430290
49. Girgis HZ (2015) Red: an intelligent, rapid, accurate tool for detecting repeats de-novo on the genomic scale. BMC Bioinformatics 16:227
50. Schaeffer CE, Figueroa ND, Liu X, et al (2016) phRAIDER: Pattern-Hunter based Rapid Ab Initio Detection of Elementary Repeats. Bioinformatics 32:i209–i215
51. Gu W, Castoe TA, Hedges DJ, et al (2008) Identification of repeat structure in large genomes using repeat probability clouds. Anal Biochem 380:77–83
52. Bao Z and Eddy SR (2002) Automated de novo identification of repeat sequence families in sequenced genomes. Genome Res 12:1269–1276
53. Price AL, Jones NC, and Pevzner PA (2005) De novo identification of repeat families in large genomes. Bioinformatics 21 Suppl 1:i351–8

54. Flynn JM, Hubley R, Goubert C, et al (2020) RepeatModeler2 for automated genomic discovery of transposable element families. Proc Natl Acad Sci U S A 117:9451–9457
55. Camacho C, Coulouris G, Avagyan V, et al (2009) BLAST : architecture and applications. BMC Bioinformatics 10:421
56. Ou S, Su W, Liao Y, et al (2019) Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline. Genome Biol 20:275
57. Saha S, Bridges S, Magbanua ZV, et al (2008) Empirical comparison of ab initio repeat finding programs. Nucleic Acids Res 36:2284–2294
58. Xu Z and Wang H (2007) LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. Nucleic Acids Res 35:W265–8
59. Ou S and Jiang N (2019) LTR_FINDER_parallel: parallelization of LTR_FINDER enabling rapid identification of long terminal repeat retrotransposons. Mob DNA 10:48
60. Ou S and Jiang N (2018) LTR_retriever: A Highly Accurate and Sensitive Program for Identification of Long Terminal Repeat Retrotransposons. Plant Physiol 176:1410–1422
61. Ellinghaus D, Kurtz S, and Willhoeft U (2008) LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons. BMC Bioinformatics 9:18
62. Han Y and Wessler SR (2010) MITE-Hunter: a program for discovering miniature inverted-repeat transposable elements from genomic sequences. Nucleic Acids Res 38:e199
63. Ye C, Ji G, and Liang C (2016) detectMITE: A novel approach to detect miniature inverted repeat transposable elements in genomes. Sci Rep 6:19688
64. Xiong W, He L, Lai J, et al (2014) HelitronScanner uncovers a large overlooked cache of Helitron transposons in many plant genomes. Proc Natl Acad Sci U S A 111:10263–10268
65. Abrusán G, Grundmann N, DeMester L, et al (2009) TEclass--a tool for automated classification of unknown eukaryotic transposable elements. Bioinformatics 25:1329–1330
66. Feschotte C, Keswani U, Ranganathan N, et al (2009) Exploring repetitive DNA landscapes using REPCLASS, a tool that automates the classification of transposable elements in eukaryotic genomes. Genome Biol Evol 1:205–220
67. Hoede C, Arnoux S, Moisset M, et al (2014) PASTEC: An Automatic Transposable Element Classification Tool. PLoS One 9:e91929
68. Yan H, Bombarely A, and Li S (2020) DeepTE: a computational method for de novo classification of transposons with convolutional neural network. Bioinformatics 36:4269–4275
69. Cruz MHP da, Cruz MHP da, Domingues DS, et al (2021) TERL: classification of transposable elements by convolutional neural networks. Brief Bioinform 22:bbaa185
70. Flutre T, Duprat E, Feuillet C, et al (2011) Considering transposable element diversification in de novo annotation approaches. PLoS One 6:e16526
71. Goerner-Potvin P and Bourque G (2018) Computational tools to unmask transposable elements. Nat Rev Genet 19:688–704
72. Satovic E (2020) Tools and databases for solving problems in detection and identification of repetitive DNA sequences. Period Biol 121-122:7–14
73. Volders P-J, Anckaert J, Verheggen K, et al (2019) LNCipedia 5: towards a reference set of human long non-coding RNAs. Nucleic Acids Res 47:D135–D139
74. Bao W, Kojima KK, and Kohany O (2015) Repbase Update, a database of repetitive elements in eukaryotic genomes. Mob DNA 6:11
75. Kielbasa SM, Wan R, Sato K, et al (2011) Adaptive seeds tame genomic sequence comparison. Genome Res. 21:487–493
76. Li H and Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25:1754–1760
77. Li H and Durbin R (2010) Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics 26:589–595
78. Langmead B, Trapnell C, Pop M, et al (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. 10:R25
79. Langmead B and Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. Nat Methods 9:357–359
80. Kent WJ (2002) BLAT--the BLAST-like alignment tool. Genome Res 12:656–664
81. Dobin A, Davis CA, Schlesinger F, et al (2013) STAR: ultrafast universal RNA-seq aligner. Bioinformatics 29:15–21
82. Trapnell C, Pachter L, and Salzberg SL (2009) TopHat: discovering splice junctions with RNA-Seq. Bioinformatics 25:1105–1111
83. Kim D, Pertea G, Trapnell C, et al (2013) TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol 14:R36
84. Kim D, Langmead B, and Salzberg SL (2015) HISAT: a fast spliced aligner with low memory requirements. Nat Methods 12:357–360
85. Kim D, Paggi JM, Park C, et al (2019) Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. Nat Biotechnol 37:907–915
86. Newkirk D, Biesinger J, Chon A, et al (2011) AREM: aligning short reads from ChIP-sequencing by expectation maximization. J Comput Biol 18:1495–1505
87. Lorenz R, Bernhart SH, Höner Zu Siederdissen C, et al (2011) ViennaRNA Package 2.0. Algorithms Mol Biol 6:26
88. Sato K, Hamada M, Asai K, et al (2009) CENTROIDFOLD: a web server for RNA secondary structure prediction. Nucleic Acids Res 37:W277–80
89. Fukunaga T, Ozaki H, Terai G, et al (2014) CapR: revealing structural specificities of RNA-binding protein target recognition using CLIP-seq data. Genome Biol 15:R16
90. Hamada M, Ono Y, Kiryu H, et al (2016) Rtools: a web server for various secondary structural analyses on single RNA sequences. Nucleic Acids Res 44:W302–W307
91. Washietl S, Hofacker IL, and Stadler PF (2005) Fast and reliable prediction of noncoding RNAs. Proc Natl Acad Sci U S A 102:2454–2459
92. Höchsmann M, Töller T, Giegerich R, et al (2003) Local similarity in RNA secondary structures. Proc IEEE Comput Soc Bioinform Conf 2:159–168
93. Macke TJ, Ecker DJ, Gutell RR, et al (2001) RNAMotif, an RNA secondary structure definition and search algorithm. Nucleic Acids Res 29:4724–4735
94. Sato K, Kato Y, Hamada M, et al (2011) IPknot: fast and accurate prediction of RNA secondary structures with pseudoknots using integer programming. Bioinformatics 27:i85–93
95. Sato K, Akiyama M, and Sakakibara Y (2021) RNA secondary structure prediction using deep learning with thermodynamic integration. Nat Commun 12:941
96. Alkan F, Wenzel A, Palasca O, et al (2017) RIsearch2: suffix array-based large-scale prediction of RNA–RNA interactions and siRNA off-targets. Nucleic Acids Res 45:e60
97. Fukunaga T and Hamada M (2017) RIblast: an ultrafast RNA–RNA interaction prediction system based on a seed-and-extension approach. Bioinformatics 33:2666–2674
98. Fukunaga T and Hamada M (2018) A Novel Method for Assessing the Statistical Significance of RNA-RNA Interactions Between Two Long RNAs. J Comput Biol 25:976–986

99. Fukunaga T, Iwakiri J, Ono Y, et al (2019) LncRRIsearch: A Web Server for lncRNA-RNA Interaction Prediction Integrated With Tissue-Specific Expression and Subcellular Localization Data. Front Genet 10:462
100. Mann M, Wright PR, and Backofen R (2017) IntaRNA 2.0: enhanced and customizable prediction of RNA–RNA interactions. Nucleic Acids Res 45:W435–W439
101. Agarwal V, Bell GW, Nam J-W, et al (2015) Predicting effective microRNA target sites in mammalian mRNAs. Elife 4
102. Wu W-S, Huang W-C, Brown JS, et al (2018) pirScan: a webserver to predict piRNA targeting sites and to avoid transgene silencing in C. elegans. Nucleic Acids Res 46:W43–W48
103. Buske FA, Bauer DC, Mattick JS, et al (2012) Triplexator: detecting nucleic acid triple helices in genomic and transcriptomic data. Genome Res 22:1372–1381
104. Zhang Y, Long Y, and Kwoh CK (2020) Deep learning based DNA:RNA triplex forming potential prediction. BMC Bioinformatics 21:522
105. Kuo C-C, Hänzelmann S, Sentürk Cetin N, et al (2019) Detection of RNA-DNA binding sites in long noncoding RNAs. Nucleic Acids Res 47:e32
106. Jenjaroenpun P, Wongsurawat T, Yenamandra SP, et al (2015) QmRLFS-finder: a model, web server and stand-alone tool for prediction and analysis of R-loop forming sequences. Nucleic Acids Res 43:W527–34
107. Davis CA, Hitz BC, Sloan CA, et al (2018) The Encyclopedia of DNA elements (ENCODE): data portal update. Nucleic Acids Res 46:D794–D801
108. Zhang Y, Liu T, Meyer CA, et al (2008) Model-based analysis of ChIP-Seq (MACS). Genome Biol 9:R137
109. Uren PJ, Bahrami-Samani E, Burns SC, et al (2012) Site identification in high-throughput RNA-protein interaction data. Bioinformatics 28:3013–3020
110. Heinz S, Benner C, Spann N, et al (2010) Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. Mol Cell 38:576–589
111. Bailey TL, Boden M, Buske FA, et al (2009) MEME SUITE: tools for motif discovery and searching. Nucleic Acids Res 37:W202–8
112. Zhang Z and Xing Y (2017) CLIP-seq analysis of multi-mapped reads discovers novel functional RNA regulatory sites in the human transcriptome. Nucleic Acids Res 45:9260–9271
113. Francescatto M, Vitezic M, Heutink P, et al (2014) Brain-specific noncoding RNAs are likely to originate in repeats and may play a role in up-regulating genes in cis. Int J Biochem Cell Biol 54:331–337
114. Nielsen MM, Tehler D, Vang S, et al (2014) Identification of expressed and conserved human noncoding RNAs. RNA 20:236–251
115. Ritchie ME, Phipson B, Wu D, et al (2015) limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res 43:e47
116. Babaian A and Mager DL (2016) Endogenous retroviral promoter exaptation in human cancer. Mob DNA 7:24
117. Grabherr MG, Haas BJ, Yassour M, et al (2011) Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotechnol 29:644–652
118. Davis MP, Carrieri C, Saini HK, et al (2017) Transposon-driven transcription is a conserved feature of vertebrate spermatogenesis and transcript evolution. EMBO Rep 18:1231–1247
119. Jang HS, Shah NM, Du AY, et al (2019) Transposable elements drive widespread expression of oncogenes in human cancers. Nat Genet 51:611–617
120. St Laurent G, Shtokalo D, Dong B, et al (2013) VlincRNAs controlled by retroviral elements are a hallmark of pluripotency and cancer. Genome Biol 14:R73
121. Le Béguec C, Wucher V, Lagoutte L, et al (2018) Characterisation and functional predictions of canine long non-coding RNAs. Sci Rep 8:13444
122. Yanai I, Benjamin H, Shmoish M, et al (2005) Genome-wide midrange transcription profiles reveal expression level relationships in human tissue specification. Bioinformatics 21:650–659
123. Kadota K, Ye J, Nakai Y, et al (2006) ROKU: a novel method for identification of tissue-specific genes. BMC Bioinformatics 7:294
124. Miao B, Fu S, Lyu C, et al (2020) Tissue-specific usage of transposable element-derived promoters in mouse development. Genome Biol 21:255
125. Shao W and Wang T (2021) Transcript assembly improves expression quantification of transposable elements in single-cell RNA-seq data. Genome Res 31:88–100
126. Hamilton RS, Hartswood E, Vendra G, et al (2009) A bioinformatics search pipeline, RNA2DSearch, identifies RNA localization elements in Drosophila retrotransposons. RNA 15:200–207
127. Hofacker IL, Priwitzer B, and Stadler PF (2004) Prediction of locally stable RNA secondary structures for genome-wide surveys. Bioinformatics 20:186–190
128. Hofacker IL (2003) Vienna RNA secondary structure server. Nucleic Acids Res 31:3429–3431
129. Zeng C, Fukunaga T, and Hamada M (2018) Identification and analysis of ribosome-associated lncRNAs using ribosome profiling data. BMC Genomics 19:414
130. Cox DR (1959) The Regression Analysis of Binary Sequences. J R Stat Soc Series B Stat Methodol 21:238–238
131. Zeng C and Hamada M (2018) Identifying sequence features that drive ribosomal association for lncRNA. BMC Genomics 19:906
132. Nadel J, Athanasiadou R, Lemetre C, et al (2015) RNA:DNA hybrids in the human genome have distinctive nucleotide characteristics, chromatin composition, and transcriptional relationships. Epigenetics Chromatin 8:46
133. Iwasaki YW, Siomi MC, and Siomi H (2015) PIWI-Interacting RNA: Its Biogenesis and Functions. Annu Rev Biochem 84:405–433
134. Petri R, Brattås PL, Sharma Y, et al (2019) LINE-2 transposable elements are a source of functional human microRNAs and target sites. PLoS Genet 15:e1008036
135. Piriyapongsa J, Mariño-Ramírez L, and Jordan IK (2007) Origin and evolution of human microRNAs from transposable elements. Genetics 176:1323–1337
136. Cho J and Paszkowski J (2017) Regulation of rice root development by a retrotransposon acting as a microRNA sponge. Elife 6:e30038
137. Nguyen TC, Cao X, Yu P, et al (2016) Mapping RNA–RNA interactome and RNA structure in vivo by MARIO. Nat Commun 7:12023
138. Ziv O, Gabryelska MM, Lun ATL, et al (2018) COMRADES determines in vivo RNA structures and interactions. Nat Methods 15:785–788
139. Zhang M, Li K, Bai J, et al (2021) Optimized photochemistry enables efficient analysis of dynamic RNA structuromes and interactomes in genetic and infectious diseases. Nat Commun 12:2344
140. Lu Z, Zhang QC, Lee B, et al (2016) RNA Duplex Map in Living Cells Reveals Higher-Order Transcriptome Structure. Cell 165:1267–1279
141. Cai Z, Cao C, Ji L, et al (2020) RIC-seq for global in situ profiling of RNA–RNA spatial interactions. Nature 582:432–437
142. Gong J, Shao D, Xu K, et al (2018) RISE: a database of RNA interactome from sequencing experiments. Nucleic Acids Res 46:D194–D201
143. Iwakiri J, Terai G, and Hamada M (2017) Computational prediction of lncRNA-mRNA interactions by integrating tissue specificity in human transcriptome. Biol Direct 12:15

144. Rafiee M-R, Zagalak JA, Sidorov S, et al (2021), Chromatin-contact atlas reveals disorder-mediated protein interactions, bioRxiv http://dx.doi.org/10.1101/2020.07.13.200212
145. Deforges J, Reis RS, Jacquet P, et al (2019) Prediction of regulatory long intergenic non-coding RNAs acting in trans through base-pairing interactions. BMC Genomics 20:601
146. Bonetti A, Agostini F, Suzuki AM, et al (2020) RADICL-seq identifies general and cell type–specific principles of genome-wide RNA-chromatin interactions. Nat Commun 11:1018
147. Zeng C, Onoguchi M, and Hamada M (2021) Association analysis of repetitive elements and R-loop formation across species. Mob DNA 12:3
148. Bai X, Li F, and Zhang Z (2021), Evidences for functional trans-acting eRNA-promoter R-loops at Alu sequences, bioRxiv http://dx.doi.org/10.1101/2021.02.17.431596
149. Fullwood MJ, Liu MH, Pan YF, et al (2009) An oestrogen-receptor-alpha-bound human chromatin interactome. Nature 462:58–64
150. Li X, Zhou B, Chen L, et al (2017) GRID-seq reveals the global RNA–chromatin interactome. Nat Biotechnol 35:940–950
151. Wu W, Yan Z, Nguyen TC, et al (2019) Mapping RNA–chromatin interactions by sequencing with iMARGI. Nat Protoc 14:3243–3272
152. Xu W, Xu H, Li K, et al (2017) The R-loop is a common chromatin feature of the Arabidopsis genome. Nat Plants 3:704–714
153. Kelley DR, Hendrickson DG, Tenen D, et al (2014) Transposable elements modulate human RNA abundance and splicing via specific RNA-protein interactions. Genome Biol 15:537
154. Harrow J, Frankish A, Gonzalez JM, et al (2012) GENCODE: the reference human genome annotation for The ENCODE Project. Genome Res 22:1760–1774
155. Kelley D, CLIP-Seq peak calling, https://github.com/davek44/CLIP-Seq. Accessed 1 May 2021
156. Quinlan AR and Hall IM (2010) BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26:841–842
157. Wheeler TJ and Eddy SR (2013) nhmmer: DNA homology search with profile HMMs. Bioinformatics 29:2487–2489
158. Beckstette M, Homann R, Giegerich R, et al (2006) Fast index based algorithms and software for matching position specific scoring matrices. BMC Bioinformatics 7:389
159. Stegmaier P, Kel A, Wingender E, et al (2013) A discriminative approach for unsupervised clustering of DNA sequence motifs. PLoS Comput Biol 9:e1002958
160. Pollard KS, Hubisz MJ, Rosenbloom KR, et al (2010) Detection of nonneutral substitution rates on mammalian phylogenies. Genome Res 20:110–121
161. Polymenidou M, Lagier-Tourenne C, Hutt KR, et al (2011) Long pre-mRNA depletion and RNA missplicing contribute to neuronal vulnerability from loss of TDP-43. Nat Neurosci 14:459–468
162. Glaz J, Pozdnyakov V, and Wallenstein S (2009) Scan Statistics: Methods and Applications, Springer Science & Business Media
163. Lee H and Schatz MC (2012) Genomic dark matter: the reliability of short read mapping illustrated by the genome mappability score. Bioinformatics 28:2097–2105
164. Trapnell C, Hendrickson DG, Sauvageau M, et al (2013) Differential analysis of gene regulation at transcript resolution with RNA-seq. Nat Biotechnol 31:46–53
165. Li YE, Xiao M, Shi B, et al (2017) Identification of high-confidence RNA regulatory elements by combinatorial classification of RNA–protein binding sites. Genome Biol 18
166. Jiang M, Anderson J, Gillespie J, et al (2008) uShuffle: A useful tool for shuffling biological sequences while preserving the k-let counts. BMC Bioinformatics 9:192
167. Kirk JM, Kim SO, Inoue K, et al (2018) Functional classification of long non-coding RNAs by k-mer content. Nat Genet 50:1474–1482