Annual Review of Genomics and Human Genetics

Investigating the Potential Roles of SINEs in the Human Genome

Xiao-Ou Zhang,1,2 Henry Pratt,1 and Zhiping Weng1

1Program in Bioinformatics and Integrative Biology, University of Massachusetts Medical School, Worcester, Massachusetts 01605, USA; email: zhiping.weng@umassmed.edu
2Current affiliation: School of Life Sciences and Technology, Tongji University, Shanghai 200092, China

Keywords
SINE, Alu, MIR, IR-Alu, gene regulation

Abstract
Short interspersed nuclear elements (SINEs) are nonautonomous retrotransposons that occupy approximately 13% of the human genome. They are transcribed by RNA polymerase III and can be retrotranscribed and inserted back into the genome with the help of other autonomous retroelements. Because they are preferentially located close to or within gene-rich regions, they can regulate gene expression by various mechanisms that act at both the DNA and the RNA levels. In this review, we summarize recent findings on the involvement of SINEs in different types of gene regulation and discuss the potential regulatory functions of SINEs that are in close proximity to genes, Pol III–transcribed SINE RNAs, and embedded SINE sequences within Pol II–transcribed genes in the human genome. These discoveries illustrate how the human genome has exapted some SINEs into functional regulatory elements.
1. INTRODUCTION

Retrotransposons are repetitive elements that can amplify themselves in the host genome via a copy-and-paste mechanism (28). A subclass called autonomous retrotransposons further contains the reverse transcriptase gene, which encodes an enzyme that enables them to be reverse transcribed and integrated into the genome. Short interspersed nuclear elements (SINEs) belong to another category, nonautonomous retrotransposons, which do not encode a functional reverse transcriptase and thus rely on other mobile elements for their retrotranscription. The initial sequencing and analysis of the human genome revealed that approximately 1.8 million copies of SINEs are scattered throughout the genome, accounting for 13% of our total DNA (73) (Figure 1a). Due to the accumulation of sequence mutations and the complicated retrotranscription machinery, only some young SINEs still have the potential to be retrotranscribed (84) and inserted into modern human genomes (69). The remaining inactive SINE elements, which have lost this ability, may be considered a sort of genomic dark matter, whose regulatory function we have not fully elucidated.

In the human genome, Alu elements and mammalian-wide interspersed repeat (MIR) elements are the two most abundant retrotransposon families of SINEs, with approximately 1.2 million and 0.6 million copies, respectively (Figure 1a). The name Alu is derived from the restriction

**Figure 1**

SINEs are distributed in a nonrandom fashion throughout the human genome. (a) The percentages of the human genome accounted for by different transposon elements. SINEs account for approximately 13% of the human genome sequence, with the two most abundant SINE families, Alu and MIR elements, accounting for 10% and 3%, respectively. (b) The canonical Alu element consists of two 7SL-derived arms, with internal Pol III promoter elements (A-box and B-box) in the left arm. The canonical MIR element consists of a tRNA-derived left arm containing Pol III promoter elements, a 70-base-pair conserved central SINE sequence, and a LINE-derived right arm. (c) The distribution of Alu and MIR elements in different genomic regions based on GENCODE annotations. Abbreviations: CDS, coding sequence; LINE, long interspersed nuclear element; LTR, long terminal repeat; Pol, polymerase; SINE, short interspersed nuclear element; UTR, untranslated region.
endonuclease from the gram-positive bacteria *Arthrobacter luteus* (*Alu I*) (104) which was used to characterize the first discovered *Alu* element in the human genome (31). *Alu* elements are approximately 280 base pairs long and derived from the head-to-tail fusion of two distinct 7SL monomers, which evolved into the left and right *Alu* arms (Figure 1b). The left arm has bipartite promoter elements (A-box and B-box) (91) (Figure 1b), bound by the RNA polymerase (Pol) III transcription factor TFIIIC, which in turn initiates Pol III transcription of the *Alu* element (88). *MIR* elements are one of the most ancient retrotransposon families (108). They are 260 base pairs in full length and consist of a tRNA-derived left arm, a 70-base-pair conserved central SINE sequence, and a long interspersed nuclear element (LINE)–derived right arm (108) (Figure 1b). Like *Alu* elements, the left arm of *MIR* elements contains an A-box and a B-box (Figure 1b), which are internal promoter elements that allow them to be transcribed by Pol III (19). Besides being transcribed from their own Pol III promoters to generate primary RNA transcripts, both *Alu* and *MIR* elements are present within the bodies of tens of thousands of Pol II–transcribed genes, leading to widespread transcription of bystander SINE RNAs by Pol II.

Given the high sequence similarity among SINE families, SINE elements form family-specific secondary structures regardless of whether they are transcribed by Pol III as independent RNAs or by Pol II as embedded sequences within genes (2, 107). Some genes even contain multiple *Alu* elements arranged in such a way that they can form inverted, repeated *Alu* pairs (IRA*lus*) that form special double-stranded RNA (dsRNA) structures (36) when transcribed. These secondary structures can bind trans-factors that play important roles in gene expression and regulation, and aberrant accumulation of these structural SINE RNAs can lead to human disorders, as discussed in detail below.

The intrinsic sequence and structural features of SINEs give them the potential to regulate gene expression and other biological processes in multiple ways. Although they have historically been thought of as junk DNA, accumulating lines of evidence show that SINEs not only provide new regulatory *cis*-elements to nearby genes but also become effectors of gene expression via co- or posttranscriptional regulation. In this review, we summarize emerging evidence revealing SINEs as a genome-wide source of regulatory elements, discuss the regulatory roles of SINEs in altering gene expression and influencing chromosomal structure, and explore the potential impacts of these findings on human health.

2. NONRANDOM DISTRIBUTION OF SINES

SINES are encountered throughout the human genome, but their distribution is not random, hinting at their functional and regulatory potential. Early cytogenetic analyses showed that *Alu* elements are enriched in transcriptionally active regions of the genome (70, 86), and the first draft of the human genome sequence supported this, identifying higher *Alu* densities in gene-dense, GC-rich genomic domains (73). More than 75% of human genes contain at least one *Alu* element, and *Alu* density is highly correlated with GC content, gene density, and intron density (47). Indeed, using GENCODE gene annotations (42), we found that 60.8% of *Alu* elements and 61.5% of *MIR* elements are within introns, while only 38.6% of *Alu* elements and 37.7% of *MIR* elements are in intergenic regions (Figure 1c). Although very few SINES have been inserted into coding exons, 3' untranslated regions (UTRs) carry thousands of SINES (Figure 1c).

Interestingly, however, newly inserted *Alu* elements do not show the same proclivity for gene-rich regions as established *Alu* elements do (30, 62), which suggests that the observed bias for gene-rich regions is not caused by insertion-site preference. Instead, it is likely that differential selection pressures operate on insertions at different genomic locations, with *Alu* elements inserted in gene-rich regions being tolerated and intergenic *Alu* elements in gene-poor regions being subject to
strong negative selection (30, 62), leading to their removal during evolution through nonallelic homologous recombination (1, 62).

SINE distribution is also biased toward different categories of genes. One interesting observation is that Alu elements are more common in regions of high-expression and housekeeping genes than in regions of low-expression and tissue-specific genes (37). Similarly, analyses of genes on human chromosomes 21 and 22 revealed that Alu elements prefer to be located in genes enriched for certain functional categories, such as metabolism, transport, and signaling, but depleted in categories such as structural proteins and information pathway components (46). The nonrandom distribution of SINES gives them the potential to be involved in biologically significant regulation of gene expression.

3. TRANSCRIPTION REGULATION BY SINE-DERIVED CIS-ELEMENTS

One possible functional role for SINE elements is to regulate transcription by acting as noncoding DNA regulatory elements, such as promoters, enhancers, or insulators. Indeed, SINES naturally possess cis-response elements, and their insertion into the genome thus leads to dispersion of potential transcription factor binding sites. In particular, the consensus sequence of Alu elements contains response elements to nuclear receptor hormones (9), calcium (81), and other effectors (74, 117). Meanwhile, SINES can also acquire regulatory potential by accumulating random mutations; for example, Alu elements are enriched in CpG dinucleotides, and the deamination of methylated CpGs in Alu elements can cause them to acquire functional binding sites for p53 and c-MYC (123, 124). Intrinsic and newly acquired binding motifs and response elements facilitate the exploitation of SINES to provide a wide variety of regulatory elements to the human genome, including promoters, enhancers, and insulators, and SINE-derived sequences can even exert higher-order influences on gene regulation by changing chromatin structure across the genome.

3.1. SINE-Derived Promoters

A comprehensive survey of experimentally characterized human promoters showed that approximately 25% of analyzed promoter regions contained a transposon-derived sequence, with SINES as the predominant transposon class (61) (Figure 2a). Specifically, SINE-driven promoters play important roles in early development (98) and some terminally differentiated tissues (93), as well as in initiating the transcription of genes involved in immunity or response to external stimuli (100, 116). Recent evidence also suggests that SINES are significant contributors to the birth of vertebrate long noncoding RNAs (lncRNAs) (26, 64).

This function is directly relevant to human health. SINES are epigenetically silenced in somatic tissues but can be epigenetically reactivated as cryptic promoters to drive oncogene expression in cancers in a process known as onco-exaptation. For example, an AluJb element located 20 kb upstream of the canonical promoter of LIN28B can function as an alternative promoter to drive the expression of LIN28B transcripts in a substantial number of tumors, especially lung cancers (58). Luciferase assays using various mutations of this AluJb element show that it contains all the necessary sequences for strong promoter activity, having accumulated mutations to acquire novel binding sites for the transcription factors CEBPD, SP1, SP4, and YY1, which in turn recruit Pol II (58). LIN28B proteins repress let-7 microRNAs to upregulate oncogenesis-related genes such as MYC and RAS; lung cancers in particular predominantly transcribe LIN28B beginning from this alternative AluJb promoter rather than the canonical promoter (58). Western blots show that the isoform derived from the AluJb-LIN28B onco-exaptation event encodes a novel protein with 22 extra amino acids at the N terminus, but, interestingly, the AluJb-LIN28B protein retains the
Roles of SINEs in the Human Genome

5.5

Review in Advance first posted on April 1, 2021. (Changes may still occur before final publication.)
SINEs can regulate gene expression by acting as noncoding DNA elements or as Pol III–transcribed RNAs. (a) SINEs can function as promoters to produce alternative isoforms for Pol II–transcribed genes. (b) In response to serum deprivation, TFIIIC binds to ADNP-bound Alu elements, which leads to H3K18ac and rewired 3D genome architecture through CTCF looping, sustaining the expression levels of cell cycle genes. (c) Pol III–transcribed SINEs can be identified by various high-throughput sequencing assays. (d) Pol III–transcribed Alu RNAs and Alu RNP s affect the assembly of the preinitiation complex to regulate gene expression at the translational level. (e) Aberrant accumulation of Pol III–transcribed Alu RNAs in DICER1-deficient cells leads to the activation of the NLRP3 inflammasome and MyD88-dependent cell death. Abbreviations: ChIP-seq, chromatin immunoprecipitation followed by sequencing; m7G, 7-methylguanosine; Pol, polymerase; RAMPAGE, RNA annotation and mapping of promoters for the analysis of gene expression; RNA-seq, RNA sequencing; RNP, ribonucleoprotein particle; SINE, short interspersed nuclear element.

normal functions of the LIN28B protein (58); thus, it is not clear whether this structural difference has functional significance.

3.2. SINEs Regulate Gene Expression via Enhancement or Chromatin Looping

SINEs have also been suggested to evolve into new enhancers or insulators. Intergenic SINEs that are relatively close to transcription start sites are enriched in epigenetic enhancer marks such as the histone modifications H3K4me1, H3K4me2, and H3K27ac in a tissue-specific manner (17, 110, 128); these putative SINE enhancers may be brought spatially close to nearby promoters through 3D chromatin interactions, thus potentially activating gene expression (17, 110, 128). In addition, a bioinformatic screen using chromatin immunoprecipitation followed by sequencing (ChIP-seq) and chromatin interaction analysis by paired-end tag sequencing (ChIA-PET) data in human CD4+ T cells identified thousands of MIR elements that might function as insulators, and three of them were experimentally validated using in vitro and in vivo enhancer-blocking assays (119). These putative MIR insulators are enriched near genes of the T cell receptor pathway and reside at T cell–specific boundaries between repressive and active chromatin (119). Interestingly, unlike typical insulators that are bound by CTCF, the MIR insulators are depleted of CTCF binding but enriched in the binding of the Pol III factor TFIIIC, suggesting that they rely on a CTCF-independent mechanism to establish chromatin and regulatory domains (119). Indeed, TFIIIC, which specifically recognizes the A-box and B-box in the internal Pol III promoter of SINEs, can assemble a functional insulator complex mediating long-range chromatin looping between SINEs and nearby transcription start sites to regulate gene expression by coordinating with other insulator proteins such as CTCF and cohesion (115). For example, upon serum deprivation (Figure 2b), TFIIIC can harness a subset of Alu elements that are premarked by activity-dependent neuroprotector homeobox protein (ADNP) to establish long-distance contacts with CTCF bound to promoters of cell cycle genes. TFIIIC can subsequently acetylate H3K18 to keep these genes poised for a rapid increase in transcription in response to serum reexposure (40).

4. POL III–TRANSCRIBED SINES AND POTENTIAL FUNCTIONS

A second functional possibility is that the RNAs produced when SINEs are transcribed by Pol III are functional. Although SINEs have Pol III–specific promoter elements, their Pol III–driven expression levels are very low under normal cellular conditions (29, 95). Indeed, the transcription of SINEs by Pol III is thought to be silenced by the host’s epigenetic machinery via DNA methylation and histone methylation (60, 118); however, the expression of primary SINE RNAs can significantly increase under some adverse environments, including heat shock (80), viral infection (90), and cancer progression (112), when silencing factors are removed.
4.1. Tissue-Specific Expression of Pol III–Transcribed Alu Elements

With a functional Pol III promoter in the left arm, Alu elements can be transcribed as independent transcripts that, given their ubiquity throughout the genome, have significant potential to regulate gene expression and cellular functions (125). Various high-throughput sequencing technologies have been applied to identify such transcripts, including cap analysis gene expression (CAGE) (39), RNA sequencing (RNA-seq) (29), or ChIP-seq of Pol III factors (85, 87) (Figure 2c). Careful inspection of the results confirms that Pol III–transcribed Alu RNAs are widely expressed, with the profile of expressed Alu elements varying by tissue and cell type (29). However, several challenges limit our ability to detect transcriptionally active Alu elements. First, Alu elements evolved from the 7SL RNA, and the 1.2 million copies of Alu elements in the human genome have a highly similar sequence; thus, it is challenging to assign short sequencing reads to the bodies of individual Alu elements. Furthermore, more than 40% of Alu elements are distributed in the bodies of Pol II–transcribed genes, and weakly expressed Pol III–transcribed Alu RNAs can be difficult to detect among much stronger Pol II transcription signals.

The development of the RNA annotation and mapping of promoters for the analysis of gene expression (RAMPAGE) assay (10) and the completion of phase 3 of the Encyclopedia of DNA Elements (ENCODE) project (38) provided an opportunity to comprehensively profile the expression pattern of Pol III–transcribed Alu RNAs in a collection of 155 biosamples covering diverse tissue and cell types. RAMPAGE is a 5′-complete cDNA sequencing assay that captures transcription start sites at single-nucleotide resolution and provides transcript connectivity via paired-end sequencing (11); it can thus efficiently address the aforementioned challenges (Figure 2c). An atlas of Pol III–transcribed Alu RNAs built using RAMPAGE data reveals that 17,249 Alu elements are expressed in at least one of the 155 ENCODE biosamples, accounting for 1.44% of the 1.2 million Alu elements annotated in the human genome (128). The expression of Pol III–transcribed Alu elements shows high tissue and cell type specificity, with biosamples in related tissues clustering together by their Alu expression profiles (128). Younger Alu elements are significantly less expressed than older Alu elements, suggesting that humans are highly effective in suppressing the Alu elements that may still have the capability to be retrotranscribed and inserted into the genome (128). Like Alu elements in general, Pol III–transcribed Alu elements are enriched near genes. Furthermore, Pol III–transcribed Alu elements are enriched near genes that are expressed in a tissue-specific manner and exert tissue-specific functions (128). Comparison of genomic and epigenetic features at expressed and unexpressed Alu elements shows that proximity to Pol II genes, increased chromatin accessibility, and the presence of active histone modifications are characteristic features of cell type–specific Pol III Alu transcription, and integrated analyses with self-transcribing active regulatory region sequencing (STARR-seq), systematic high-resolution activation and repression profiling with reporter tiling using massively parallel reporter assays (Sharpr-MPRA), and CRISPR/Cas9 quantitative trait locus (crisprQTL) data further support that expressed Alu elements may, in some instances, function as cell type–specific enhancers for nearby protein-coding genes (128).

4.2. Structural Pol III–Transcribed Alu RNAs Regulate Protein Translation

Besides functioning as cell type–specific enhancers when transcribed by Pol III (128), primary Alu RNAs can assemble Alu ribonucleoprotein particles (RNP)s and act as trans-regulatory factors to modulate protein translation. A typical Alu RNA is a fusion of two nonequivalent 7SL-derived arms, which resemble the 7SL RNA component of the signal recognition particle (SRP) in terms of sequence and secondary structure. Composed of six SRP proteins and a 7SL RNA, SRP is...
an essential ribonucleoprotein complex responsible for cotranslational delivery of membrane and secretory proteins to the endoplasmic reticulum in human cells (5). Alu RNAs are sufficiently similar to the SRP 7SL RNA and, as a result, bind the SRP protein SRP9/14 subunit of SRP in vitro (14) and in vivo (20), with the left arm exhibiting a higher affinity (14). The SRP9/14 subunit is required for SRP to arrest elongation to ensure efficient protein translocation into the endoplasmic reticulum (72).

In vitro experiments show that purified synthetic Alu-SRP9/14 RNPs interfere with translation initiation by inhibiting polysome formation (50). The canonical pathway for translation initiation proceeds through the sequential formation of several initiation complexes (reviewed in 57). The ternary complex (formed by eIF2, GTP, and Met-tRNAMet\(^{\text{Met}}\)) and the initiation factors (eIF1, eIF1A, eIF3, and likely eIF5) first bind the 40S ribosomal subunit to yield the 43S preinitiation complex, which is then loaded onto the 5′ UTR of an mRNA to scan for the start codon. Once the start codon is positioned at the P-site, the 48S initiation complex is assembled, switching the scanning 43S preinitiation complex to a closed conformation; this 48S complex then combines with the 60S ribosomal subunit to form the 80S initiation complex, which then enters the elongation cycle. Although SRP is not normally involved in translation initiation, Alu-SRP9/14 RNPs inhibit translation by interacting directly with the 40S ribosomal subunit, which prevents the recruitment of the 43S preinitiation complex to mRNA, possibly due to the direct or indirect blockage of the mRNA from entering into the 43S complex (56) (Figure 2d). Alu binding enhances the delivery of the inhibitive SRP9/14 to the 40S subunit, but Alu binding is not required for the continued association between SRP9/14 and the 40S subunit—the Alu can dissociate and recruit another SRP9/14 to another 40S subunit to inhibit the formation of another 48S complex while leaving the first SRP9/14 bound (13, 56) (Figure 2d).

4.3. Accumulation of Pol III–Transcribed Alu RNAs in Stress Response

Many studies have shown that the expression of Pol III–transcribed Alu RNAs dramatically increases when cells encounter certain external stresses (80, 90, 112). This increase could represent a nonfunctional consequence of a global change in genome transcription in response to the stress or could act as a regulatory effector to overcome stress conditions en route to recovery. During heat shock response, accumulated Pol III–transcribed Alu RNAs can bind directly to Pol II to block the assembly of the transcription initiation complex, resulting in global transcription repression (80). The function of Alu RNAs as transcriptional repressors is dependent on its secondary structures, as shown by in vitro transcription combined with deletion analysis, which identified two structured domains in the right arm and middle linker of Alu RNAs (80). On the other hand, deficiency in the degradation of the Alu transcript can lead to aberrant accumulation of Pol III–transcribed Alu RNAs during the stress response, causing Alu toxicity. One well-studied case is Pol III–transcribed Alu RNA accumulation in the retinal pigment epithelium (RPE) of human eyes with geographic atrophy (63), which results in the late-stage age-related macular degeneration that is a leading cause of blindness. Geographic atrophy is characterized by scattered or confluent areas of degeneration of RPE cells and the light-sensing retinal photoreceptors that overlie the RPE cells (7). In geographic atrophy patients, pathological decrease in the protein level of the DICER1 ribonuclease, which cleaves dsRNAs and pre-microRNAs into short double-stranded RNA fragments (109), leads to deficient degradation and aberrant accumulation of Pol III–transcribed Alu RNAs in RPE cells (63) (Figure 2e), which subsequently activates the NLRP3 inflammasome (111), leading to the secretion of interleukin-18 (IL-18) and MyD88-dependent RPE cell death (113) (Figure 2e).
5. CO- AND POSTTRANSCRIPTIONAL REGULATION BY SINES EMBEDDED IN POL II–TRANSCRIBED GENES

A third functional possibility is that SINEs transcribed within a Pol II–transcribed RNA can act as RNA regulatory elements, influencing the fate of the RNA in which they reside. SINEs occur frequently in introns and UTRs of human genes, leading to the pervasive transcription of embedded SINE sequences as a part of Pol II–transcribed RNAs. Accumulating lines of evidence support roles for these SINE elements in a number of co- and posttranscriptional processes that influence gene expression. In this section, we first review the role of SINEs in splicing, nuclear mRNA retention, and Staufen (STAU)–mediated mRNA decay, and then discuss the implications of competition between these functions in regulating higher-order biological processes, using the innate immune response and its potential applications in cancer therapies as two examples.

5.1. Complementary SINE-Mediated Exon Circularization and Alternative Back-Splicing

With regard to splicing, embedded SINEs are perhaps most influential in regulating a process called back-splicing (exon circularization), in which a downstream 5′ splice site (splice donor) joins with an upstream 3′ splice site (splice acceptor), in the reverse order of regular splicing, across one or more exons to form covalently circularized RNA transcripts or circular RNAs (circRNAs). Many protein-coding genes in the human genome can undergo this process; the reaction is inefficiently catalyzed by the spliceosome, and thus circRNAs are generally expressed at low levels, but they have important effects, modulating gene expression via transcription, splicing, and sequestering or scaffolding macromolecules (21). Emerging lines of evidence also show that circRNAs play an important role in regulating immune responses and cell proliferation (reviewed in 21).

SINE elements can influence back-splicing of exons when they are present in adjacent flanking introns. High-throughput sequencing technology has identified tens of thousands of circRNAs in the human genome (59, 83, 129), and large-scale analyses show that back-spliced exons are generally flanked by long introns, wherein inverted complementary Alu elements can form IR-Alus to bring distal splice sites into close proximity and facilitate back-splicing (59, 76, 129) (Figure 3a). This is true even for the first well-characterized mammalian circRNA, circSRY. First described in the 1990s, circSRY was from the mouse Sry gene (18, 34), which possesses very long reverse complementary sequences flanking its exons (34). IR-Alus can also form between inverted complementary Alu elements within the same flanking intron to facilitate canonical splicing for linear RNA formation (Figure 3b), and competition between the Alu elements in the same flanking intron and the Alu elements in the two flanking introns can modulate the relative efficiency of back-splicing and splicing (129) (Figure 3b). Thus, the SINEs that lie in the introns of human genes, which account for 61% of SINEs (Figure 1c), greatly increase the complexity of splicing patterns and splicing regulation in the human genome. Indeed, a comparison of human and mouse circRNA expression showed that the majority (~85%) of circRNAs are specifically expressed in humans but not mice—likely owing to the strong pairing between intronic human-specific SINEs, especially Alu elements (33).

Competition between SINEs adds an additional layer of complexity in RNAs containing two or more exons; in such RNAs, a given locus can produce multiple circRNAs by alternative back-splicing (ABS) (127) (Figure 3c). There are two types of ABS, each using a different set of downstream 5′ splice donor and upstream 3′ splice acceptor. In alternative 5′ back-splicing (A5BS), two or more downstream 5′ splice donors are alternatively back-spliced to the same upstream 3′ splice acceptor. In alternative 3′ back-splicing (A3BS), two or more upstream 3′ splice acceptors are alternatively back-spliced to the same downstream 5′ splice donor. Large-scale transcriptome
**5.10** Zhang • Pratt • Weng

![Diagram showing molecular interactions](image)

**Caption appears on following page**
analyses in 90 human tissue samples revealed that ABS pervasively occurs during the biogenesis of circRNAs, with 84% of circRNAs being ABS circRNAs (126). In the flanking introns of more than 70% of ABS circRNAs, IR\textit{Alu}s can pair across different sets of back-spliced exons to regulate ABS events (127). The number of\textit{Alu} elements in flanking introns and their pairing capacities significantly affect the complexity of ABS, defined as the number of different circRNAs in an ABS event; specifically, longer flanking introns can accommodate more\textit{Alu} elements and intensify competition between pairings across different sets of back-spliced exons (126) (\textbf{Figure 3c}). Thus, A5BS events with higher complexity generally have longer upstream flanking introns, which in turn contain more\textit{Alu} elements, than A3BS events with lower complexity. In the reciprocal case, more complex A3BS events tend to have longer downstream flanking exons, which contain more\textit{Alu} elements, than A3BS events with lower complexity. Similarly, the most abundant (predominant) circRNAs in A5BS events have longer downstream flanking introns than the remaining circRNAs in the same A5BS events, while the predominant circRNAs in A3BS events have longer upstream flanking introns than the remaining circRNAs in the same A5BS events. The\textit{Alu} elements in these long introns help the predominant circRNAs to outcompete other circRNAs in the same ABS event. Meanwhile, IR\textit{Alu}s with higher pairing capacity in flanking introns can also assist the corresponding circRNAs to outcompete other circRNAs to become the predominant circRNA in ABS events (126) (\textbf{Figure 3e}).

Despite having the same IR\textit{Alu}s in the flanking introns, circRNAs and ABS events exhibit highly tissue-specific or cell type–specific patterns (126, 127), suggesting the involvement of other\textit{trans}-factors. Because IR\textit{Alu}s can form stable double-stranded RNA structures, RNA-binding proteins with a double-stranded RNA-binding domain (dsRBD) can bind and stabilize the transiently formed intronic IR\textit{Alu} flanking back-spliced exons to facilitate back-splicing. Several examples are known. One well-characterized example is the interleukin enhancer-binding factor 3 (ILF3) gene, which encodes two proteins, nuclear factor 90 (NF90) and nuclear factor 110 (NF110) (94); these proteins contain two dsRBDs, each of which can directly bind IR\textit{Alu}s to promote circRNA formation (75). Knockout of NF90 or NF110 leads to a global reduction in the levels of nascent circRNAs, which can be rescued with wild-type NF90 but not dsRBD-mutated NF90 (75). Another example is DHX9, a nuclear RNA helicase with both a dsRBD and an RNA helicase domain, which can dampen circRNA formation by destabilizing the pairing of IR\textit{Alu}s (6). When binding to the intronic IR\textit{Alu}s, DHX9 will unwind RNA pairs flanking back-spliced exons using its RNA helicase activity, making back-splicing less efficient (6); consistent with this mechanism, the depletion of DHX9 increases circRNA production from the IR\textit{Alu}-rich survival motor neuron (SMN) genes (89). A third example is adenosine deaminase acting on RNA 1 (ADAR1), a dsRNA-specific adenosine deaminase, which can also inhibit back-splicing (55, 102). ADAR1 converts adenosines (A) in dsRNAs to inosines (I), a process known as A-to-I editing. When ADAR1 performs RNA editing on IR\textit{Alu}s (8, 65), it impairs the pairing stability of IR\textit{Alu}s and inhibits the back-splicing of neighboring circRNAs (55, 102). The relatively long flanking introns of circRNAs and their abundantly distributed\textit{Alu} elements offer many potential...
binding sites for RNA-binding proteins to shape the tissue-specific ABS patterns. Because these RNA-binding proteins all bind to IR\textit{Alu} flanking back-spliced exons, it remains unclear whether and (if so) how they compete or coordinate with each other to regulate the ABS of circRNAs.

### 5.2. SINE-Mediated Nuclear Retention

After splicing is complete, gene expression can be regulated by the retention of mature mRNAs in the nucleus (48), a process that is mediated by SINES embedded in genes. One well-studied nuclear retention pathway is mediated by IR\textit{Alu} and subnuclear bodies called paraspeckles. In the nucleus, IR\textit{Alu} is the 3′ UTR of mRNAs are frequently A-to-I edited by the adenosine deaminase ADAR1 (8, 65). mRNAs with structured and edited IR\textit{Alu} are bound by the protein NONO, which concentrates in paraspeckles (41) and prevents the export of the mRNAs to the cytoplasm (22, 23) (Figure 3d). This process can be further modulated by posttranslational modifications of NONO, which affects its affinity for IR\textit{Alu}. For example, coactivator-associated arginine methyltransferase 1 (CARM1) can methylate the coiled-coil domain of NONO, which decreases its affinity for 3′ UTR IR\textit{Alu} and promotes nuclear export of IR\textit{Alu}-containing mRNAs (52). Besides NONO, the nuclear-retained noncoding RNA \textit{NEAT1}, which is essential to paraspeckle integrity (79, 103), is also required for effective retention of edited IR\textit{Alu}-containing mRNAs in the nucleus (22).

IR\textit{Alu} can also occur within, and regulate the subcellular localization of, lncRNAs. Human \textit{lincRNA-p21} is an IR\textit{Alu}-containing lncRNA that can promote apoptosis through a feedback mechanism that enhances \textit{TP53} transcriptional activity in the nucleus (120). It can also be exported to the cytoplasm, where it represses the translation of specific target genes through activation of the RNA-induced silencing complex (RISC) and induces glycolysis under hypoxic conditions (121, 122). The dynamic subcellular localization of human \textit{lincRNA-p21}, two primate-conserved inverted-repeat \textit{Alu} elements fold into the IR\textit{Alu} structure, mediating human \textit{lincRNA-p21} colocalizing with paraspeckles during the course of the stress response (25).

Such paraspeckle-mediated nuclear retention provides a regulatory mechanism that prevents inappropriate translation of ADAR-edited RNAs and retains nuclear lncRNAs in the nucleus to exert biological functions.

Another recently discovered SINE-mediated mechanism for nuclear retention involves a 42-nucleotide-long fragment named SINE-derived nuclear RNA localization (SIRLOIN) (78), which is present in some \textit{Alu} elements that are inserted in exons in the antisense orientation. SIRLOIN elements contain three stretches of at least six pyrimidines (C or T), with two stretches matching the consensus RCCTCCC (where R denotes A or G) (78). A survey of ENCODE enhanced cross-linking and immunoprecipitation (eCLIP) data sets showed that binding sites of heterogeneous ribonucleoprotein K (HNRNPK) are specifically enriched in SIRLOINs. Knockdown of the \textit{HNRNPK} gene by short interfering RNAs (siRNAs) resulted in significantly lower levels of SIRLOIN-containing transcripts in the nucleus, suggesting that the nuclear retention of these SIRLOIN-containing transcripts was mediated by HNRNPK (78). Although the detailed mechanism by which HNRNPK affects nuclear retention is not yet fully understood, it is expected that overexpression or mutation of \textit{HNRNPK} would affect the relative localization of SIRLOIN-containing RNAs in the nucleus and cytoplasm, which might lead to biological effects and pathogenic conditions.

### 5.3. 3′ UTR SINEs and STAU-Mediated mRNA Decay

After export from the nucleus, \textit{Alu} elements can also influence mRNA expression by mediating STAU-mediated mRNA decay (SMD). SMD is an mRNA degradation process that occurs in the
cytoplasm of mammalian cells and plays important roles in cell motility, cell invasion, and other processes (reviewed in 92). STAU proteins possess a dsRBD that recognizes dsRNA structures in the 3′ UTR of target mRNAs (43), and the most pervasive dsRNA structures in the 3′ UTR of human genes are formed by Alu elements. dsRNA structures can occur intermolecularly between Alu sequences in different transcripts or intramolecularly between multiple IR-Alu within the same transcript; both result in dsRNAs that create binding sites for STAU1 and/or its paralog STAU2 (43, 44). Upon binding, STAU proteins recruit UPF1 to the 3′ UTR of the target mRNA to trigger mRNA decay (92) (Figure 3d). SMD is a translation-dependent process and only targets translationally active mRNA (92). When STAU proteins bind an intermolecular Alu–Alu duplex formed between two transcripts, they will trigger SMD for both of the transcripts in the duplex only if they are both translated (44). If only one transcript is actively translated and the other is a IncRNA or a translationally inactive mRNA, only the translationally active transcript will be targeted for SMD (44).

5.4. IR-Alu and Innate Immune Response

The innate immune response is a well-studied example of biological processes that are regulated via Alu elements. Many proteins in the innate immune system initiate antiviral signaling upon recognizing foreign nucleic acids from invading pathogens, and foreign RNAs frequently form dsRNA structures, as Alu elements do. Downstream, immune response proteins exert their antiviral effects through the production of type I interferons and induction of interferon-stimulated genes (reviewed in 54); however, when these proteins recognize IR-Alu, the antiviral machinery can be triggered even in the absence of infection (66). For example, the interferon-induced dsRNA-activated protein kinase (PKR) phosphorylates eukaryotic translation initiation factor 2α (eIF2α), which leads to global translational suppression (12). Genome-wide profiling of the substrates of PKR activation using the formaldehyde cross-linking, immunoprecipitation, and sequencing (CLIP-seq) technique showed that more than 20% of sequencing reads were mapped to SINEs, mainly IR-Alu (68). Indeed, PKR can be activated in uninfected cells, especially during mitosis, by binding to IR-Alu embedded in the 3′ UTR of numerous mRNAs (67).

Competition between the various RNA regulatory roles of SINE elements has significant implications for higher-order biological processes in general. PKR competes with many other dsRNA-binding proteins for binding to IR-Alu at 3′ UTRs, including the aforementioned NONO (22, 23), STAU1 (43), and ADAR1 (8, 65); this competition can dynamically influence the metabolism of the mRNAs that contain 3′ UTR IR-Alu and thereby global translational regulation. In the nucleus, STAU1 proteins compete with NONO to bind 3′ UTR IR-Alu, which can dampen paraspeckle-mediated nuclear retention and increase mRNA export to the cytoplasm (Figure 3d). In the cytoplasm, binding of STAU1 to 3′ UTR IR-Alu precludes PKR binding, alleviating translational repression (35) (Figure 3d). During the interferon response, ADAR1 primarily edits IR-Alu in Pol II–transcribed mRNAs to destabilize the dsRNA structure of IR-Alu, preventing the activation of PKR and the downstream translational repression and cell death (27) (Figure 3e).

Ultimately, the disturbance of these Alu-mediated regulatory pathways can influence human health. One known example is mediated by the innate immune response protein MDA5, encoded by the IFIHI gene. MDA5 recognizes dsRNAs (32); upon binding dsRNAs, it undergoes filament formation along the RNA and induces interferon production via the downstream adapter mitochondrial antiviral signaling proteins (32). Activation of MDA5 signaling by IR-Alu can be a driver for the inflammatory disorder Aicardi-Goutières syndrome (3). Wild-type MDA5 cannot efficiently recognize the dsRNA structure of ADAR1-edited IR-Alu due to its limited ability to form filaments on imperfect duplexes (3) (Figure 3e); however, in patients with Aicardi-Goutières
syndrome, MDA5 has a gain-of-function mutation that reduces the sensitivity of MDA5 to duplex structural irregularities. The mutant MDA5 can thus assemble signaling-competent filaments on the dsRNA structure of ADAR1-edited IR\(\text{Alu}\)s, resulting in self-triggered downstream signaling (3). Similarly, ADAR1 deficiency can cause autoinflammatory diseases through aberrant activation of MDA5 (77, 96, 99). Consistent with this phenomenon, knockout of ADAR1 prevents A-to-I editing, and unedited IR\(\text{Alu}\)s are sufficient to trigger the activation of even wild-type MDA5 (3, 27) (Figure 3e). Considering that PKR can also modulate MDA5 activity (97), studying the crosstalk and redundancy between PKR and MDA5 as downstream effectors of ADAR1 deficiency in Aicardi-Goutières syndrome and other inflammatory disorders could be a fruitful avenue of future research (Figure 3e).

5.5. IR\(\text{Alu}\)s and Cancer Therapeutics

Recent breakthroughs in our understanding of the interplay between the immune system and cancer cells have led to a new generation of powerful therapies that mobilize the host immune system to recognize and eliminate cancer cells. Besides cancer immunotherapy (reviewed in 106), which has brought about long-term survival for a subset of cancer patients, epigenetic therapy and spliceosome-targeted therapy (STT) invoke the antiviral immune responses to kill cancer cells. The latest results reveal that dsRNAs from IR\(\text{Alu}\)s are the immunogenic agents responsible for the efficacy of both epigenetic therapy and STT.

Epigenetic therapy uses small-molecule inhibitors of DNA methyltransferases or histone deacetylases to derepress retroelements. The resulting dsRNAs from these retroelements mimic viral infection and activate the MDA5 pattern recognition receptor, which stimulates innate and adaptive immune responses against cancer cells (24, 101). Earlier studies performed RNA-seq on patient-derived colorectal cancer cells treated with 5-AZA-\(2'\)-deoxycytidine (5-AZA-CdR), a DNA methyltransferase inhibitor (DNMTi), revealing a derepression of LINEs and endogenous retroviruses but not SINEs (16, 24, 101). A recent study using an MDA5-protection assay followed by RNA-seq revealed that SINEs, in particular \(\text{Alu}\) elements, were markedly induced upon DNMTi treatment of patient-derived colorectal cancer cells (82). \(\text{Alu}\) elements constitute 88.72% of DNMTi-induced immunogenic RNA (82); they are protected by MDA5 and hence were not detected in earlier studies. Most of these \(\text{Alu}\) elements reside in intronic and intergenic regions downstream of orphan CpG islands, which are normally repressed by DNA methylation but become demethylated by DNMTi and act as sites of cryptic transcription initiation (82). Given that A-to-I editing by ADAR1 can destabilize the secondary structure of IR\(\text{Alu}\)s and prevent MDA5 activation (77) (Figure 3e), ADAR1 knockout in these cancer cells resulted in a significant increase in cytoplasmic dsRNA. Interestingly, cells treated with 5-AZA-CdR showed a sustained up-regulation of ADAR1, revealing a negative-feedback loop by ADAR1 to restrict the viral mimicry response to epigenetic therapy (82). The combination of ADAR1 knockdown and 5-AZA-CdR led to remarkable antitumor activity in mouse xenograft models of these cancer cells. These findings highlight a promising strategy by combining epigenetic therapies with ADAR1 inhibitors (82).

Cancer cells are replete with mutations in the RNA splicing machinery, and many oncogenic lesions not associated with the spliceosome also deregulate splicing globally, such as hyperactivation of the \(MYC\) oncogene (51). STTs exploit the hypersensitivity of these tumors to the splicing pathway via pharmacological and genetic perturbations of the spliceosome (45, 51). Accordingly, small-molecule spliceosomal inhibitors exhibit potent antitumor activity across a variety of spliceosome-sensitive cancers (49, 114), but the downstream mechanisms by which they reduce cancer cell fitness are not well understood. Previous studies have sought to explain the effects by studying the perturbed splicing patterns of specific individual genes (4, 53). A recent
study of triple-negative breast cancer, however, revealed a more general mechanism involving the dsRNA-activated antiviral immune response for STTs (15). Treatment using two small-molecule spliceosomal inhibitors, SD6 (71) and H3B-8800 (105), in triple-negative breast cancer cell lines elevates splicing errors, leading to cytosolic accumulation of intron-retaining transcripts (15). Given that a large proportion of SINEs in the genome are located in introns (Figure 1c), IR\textit{Alu}s formed in intron-retained mRNAs can cause a significant increase in cytoplasmic dsRNA; indeed, this is detectable by immunofluorescence staining following STT treatment (15). A forward genetic screen with a short hairpin RNA library targeting signal-transducing proteins revealed that this STT-induced dsRNA accumulation is sufficient to trigger antiviral signaling pathways, extrinsic apoptosis, and immune surveillance (15), underscoring the role that IR\textit{Alu}s play in the novel mechanism of STTs to augment cancer cell immunogenicity.

6. OUTLOOK

The ubiquity of SINEs throughout the human genome and their enrichment in gene-rich regions has led them to be inextricably involved in various mechanisms of regulating gene expression. Meanwhile, involvement in gene regulation allows SINEs to accumulate beneficial mutations and escape negative selection during evolution, driving them to evolve further to become regulatory elements that are even more functional. There are undoubtedly other regulatory functions of SINEs that remain to be deciphered. The findings reviewed here clearly demonstrate, however, that SINEs play numerous roles in the human genome and transcriptome at various phases of gene regulation, and our mechanistic understanding of SINEs can be harnessed to treat human diseases such as autoinflammatory disorders and cancer.

DISCLOSURE STATEMENT

Z.W. is a cofounder of Rgenta Therapeutics and serves on its board of directors and scientific advisory board.

ACKNOWLEDGMENTS

We thank members of the Weng laboratory for helpful comments on this article. This work was funded by National Institutes of Health grant U24-HG009446 to Z.W.

LITERATURE CITED

98. Polak P, Domany E. 2006. Alu elements contain many binding sites for transcription factors and may play a role in regulation of developmental processes. *BMC Genom.* 7:133


