UNIVERSITÀ DEGLI STUDI DI PARMA

PhD program in Biotechnology
XXVIII cycle

Retrotransposon expression profiling: Unveiling the hidden SINE transcriptome through Next-Generation Sequencing data analysis

Coordinator
Prof. Nelson Marmiroli

Tutor
Prof. Giorgio Dieci

PhD student
Davide Carnevali

Abstract

Of the ~1.7 million SINE elements in the human genome, only a tiny number are estimated to be active in transcription by RNA polymerase (Pol) III. Tracing the individual loci from which SINE transcripts originate is complicated by their highly repetitive nature. By exploiting RNA-Seq datasets and unique SINE DNA sequences, we devised a bioinformatic pipeline allowing us to identify Pol III-dependent transcripts of individual SINE elements. When applied to ENCODE transcriptomes of seven human cell lines, this search strategy identified ~1300 *Alu* loci and ~1100 MIR loci corresponding to detectable transcripts, with ~120 and ~60 respectively *Alu* and MIR loci expressed in at least three cell lines. *In vitro* transcription of selected SINEs did not reflect their *in vivo* expression properties, and required the native 5'-flanking region in addition to internal promoter. We also identified a cluster of expressed *Alu*Ya5-derived transcription units, juxtaposed to *snaR* genes on chromosome 19, formed by a promoter-containing left monomer fused to an *Alu*-unrelated downstream moiety. Autonomous Pol III transcription was also revealed for SINEs nested within Pol II-transcribed genes raising the possibility of an underlying mechanism for Pol II gene regulation by SINE transcriptional units. Moreover the application of our bioinformatic pipeline to both RNA-seq data of cells subjected to an *in vitro* pro-oncogenic stimulus and of in vivo matched tumor and non-tumor samples allowed us to detect increased *Alu* RNA expression as well as the source loci of such deregulation. The ability to investigate SINE transcriptomes at single-locus resolution will facilitate both the identification of novel biologically relevant SINE RNAs and the assessment of SINE expression alteration under pathological conditions.

Aknowledgements

I would like to begin by thanking my tutor, prof. Giorgio Dieci, whose experience, knowledge and advices have led my path in biological research field throughout the undergraduate and postgraduate studies.

I thank prof. Matteo Pellegrini of Molecular, Cell and Developmental Biology Department, UCLA who give me hospitality during my PhD and let me use their computational resources greatly speeding up our analysis, prof. Arnold J. Berk of Microbiology, Immunology, and Molecular Genetics, Jonsson Comprehensive Cancer Center, UCLA for providing RNA-seq data of the virus infected IMR90 cells and Dr Roberto Ferrari, PhD of Gene Regulation, Stem Cells and Cancer Program, Centre de Regulació Genòmica (CRG) for performing and analyzing ChIP-seq data in the same cells.

I also thank Dr Anastasia Conti, PhD for plasmid construction and *in vitro* transcription.

Index

1. Introduction

1.1. History of Transposable Elements (TEs)

Transposable elements (TEs), also known as "jumping genes" or transposons, are sequences of DNA that can move (or jump) within genomes. They were discovered in the 1940s by maize geneticist Barbara McClintock (1) who also suggested that these mysterious elements could have played a role in gene regulation determining which genes (and when) were 'turned on'. Later on, during the 1960s, Roy Britten and colleagues proposed that TEs played a role not only in gene regulation, but also in cell differentiation. They hypothesized that genome complexity involves the coordination and regulation of unrelated genes via a regulatory element, which could target specific unlinked genes, and they suggested the regulatory sequences were the previously described interspersed repeat elements (2).

These early speculations were largely dismissed by the scientific community, and for decades TEs were considered useless and referred to as 'Junk DNA'. Only recently have biologists begun to entertain the possibility that this so-called "junk" DNA might not be junk after all.

Ideed, following the sequencing of the human genome in 2001 (3), TEs have been discovered to make up to ~45% of it (being likely an underestimate, as many ancient TEs in the human genome have probably diverged beyond recognition) (Figure 1) and thanks to worldwide collaborative efforts such as ENCODE (Encyclopedia of DNA Elements), our understanding of these elements has greatly expanded. The study of TEs is an emerging field of research and it is now widely accepted that they played a crucial role in genome evolution and are involved in a variety of human diseases and regulatory processes many of which still have to be elucidated.

1.2. Classification of TEs

TEs are grouped into two classes: Class II elements or DNA transposons and Class I elements or retrotransposons.

DNA transposons, which are currently inactive in mammals, comprise about 3% of the human genome. Most of them move by a "cut-and-paste" non-replicative mechanism in which the transposon is excised from one location and reintegrated elsewhere and are therefore considered the least successful among TEs.

Class I elements comprise Long Terminal Repeat (LTR) retrotransposons and non-LTR retrotransposons and move by a "copy-and-paste" mechanism involving the reverse transcription of an RNA intermediate and insertion of its cDNA copy at a new site in the genome.

While LTR retrotransposons, such Endogenous Retrovirus (ERV), undergo a reverse transcription in virus-like particles by a complex multistep process and their cDNA is subsequently integrated in the genome, the RNA copies of non-LTR retrotransposons are directly carried back to the nucleus where they are integrated and reverse transcribed in a single step (4, 5).

Two classes compose non-LTR retrotransposons: Long Interspersed Nuclear Elements (LINEs) and Short Interspersed Nuclear Elements (SINEs).

While LTR retrotransposons and LINEs have the ability to direct their own amplification and are therefore called 'autonomous', SINEs are non-autonomous and rely on the retrotransposition machinery encoded by LINEs for their amplification.

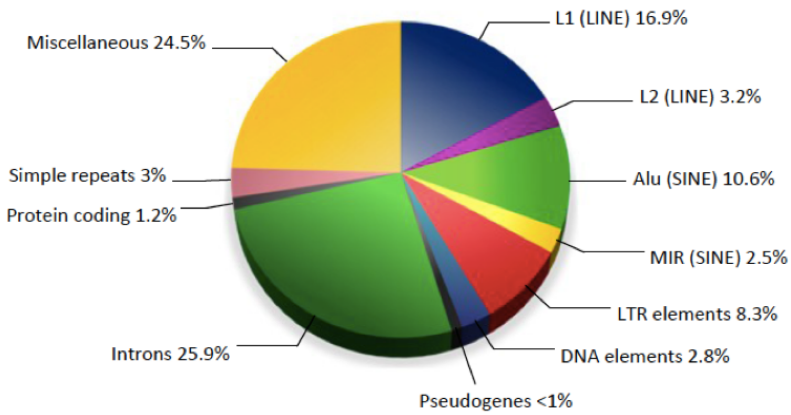


Figure 1. The principal components of the human genome. Almost 26% of the total human genome is composed of intronic sequence whereas only 1.2% encodes for protein. TEs comprise ~45% of the total genome sequence, with the most abundant elements being the LINE and SINE repeats which comprise 21.1% and 13.1% of the genome, respectively.

1.3. Long Interspersed Nuclear Elements (LINEs)

Long Interspersed Nuclear Elements are the most ancient and successful class of retrotransposons in eukaryotic genomes. They build up roughly 20% of the human genome, and are grouped in three distantly related families: LINE1 (L1), LINE2 (L2) and LINE3 (L3).

LINE1s are the youngest and most abundant LINEs in the human genome with over 500000 copies (~17% of the total genome sequence) and the only currently active with the L1Hs element (3).

LINEs are about 6-7 kb long, harbor an internal RNA Polymerase II (Pol II) promoter in their 5' UTR which is preserved following retrotransposition, and contain two open reading frames (ORFs) necessary for their amplification, thus making them autonomous retrotransposons (Figure 2). ORF1, ~1 kb in length, encodes a protein with nucleic acid binding ability that is hypothesized to bind the LINE RNA during retrotransposition (6, 7). ORF2 is ~4 kb long and encodes a protein which has both endonuclease and reverse transcriptase activity (8, 9). These proteins can also mobilize non-autonomous retrotransposons (especially SINEs), other noncoding RNA and mRNA leading to the generation of processed pseudogenes (10).

During the retrotransposition process, termed target-primed reverse transcription (TPRT) (11), reverse transcription frequently fails to proceed up to the 5' end, thus resulting in truncated, 3' enriched non-functional insertions. For this reason most of the L1 elements in the human genome are short with an average length of 900 bp for all copies. Moreover, in addition to 5' truncations, inversions and/or point mutations within the two encoded ORFs, have rendered 99.9% of human L1s inactive (12). As a result, it has been estimated that the human genome contains only ~80-100 L1Hs active elements (13). Depending upon the method used in the analysis, estimations of L1 insertions range from 1 out of 20 births in humans based on disease-causing *de novo* insertions, but approximately 1 out of 200 births based on genome comparisons (14).

Figure 2. LINE-1 (L1) element structure. The canonical L1 element harbour an internal Polymerase II promoter in its 5' UTR (blue shadow) and a polyadenilation signal (pA) preceding the oligo(dA)-rich tail (AAA). Adapted from (12)

1.4. Short Interspersed Nuclear Elements (SINEs)

SINEs are short retroelements, transcribed by RNA Polymerase III (Pol III), of size ranging between 70 and 500 bp that invaded eukaryotic genomes and with more than 1.7 x 10^6 copies build up roughly 13% of the human genome.
Because they rely on enzymatic machinery encoded by L1 to direct their retrotransposition (15), SINEs are considered non-autonomous retrotransposons. The 5' terminal parts ("heads") of all SINEs demonstrate a clear similarity with one of three types of cellular RNAs synthesized by Pol III: tRNA, 7SL RNA, or 5S rRNA, with the tRNA-derived ones being particularly abundant, except in human where 7SL-derived *Alus* are by far the most numerous (16).
Two different families belong to this class in the human genome: *Alu* and MIR (Mammalian-wide Interspersed Repeat), although there is another family of retrotransposons, SVA (SINE/VNTR/*Alu*) transcribed by Pol II, which is usually considered to be a third class of retroelements, but is sometimes classified as SINE (17). Among these families, only MIRs are currently inactive in their non-autonomous retrotranspositions process.

The RNA Polymerase III transcription machinery is devoted to the production of non-protein coding (nc) RNAs of small size, being assisted in this task by complex sets of basal and regulatory transcription factors (TFs) which vary depending on the type of promoter involved (Figure 3). SINE elements, like tRNA genes, posses a type 2 promoter which consists of two internal control regions known as A- and B-boxes forming a bipartite binding site for the multisubunit, basal transcription factor (TF) TFIIIC on the DNA. Once bound to the DNA TFIIIC recruits, on DNA upstream the transcription start site (TSS), the TFIIIB initiation complex

composed of TBP, BDP and BRF1 proteins which in turns recruits Pol III to initiate transcription.

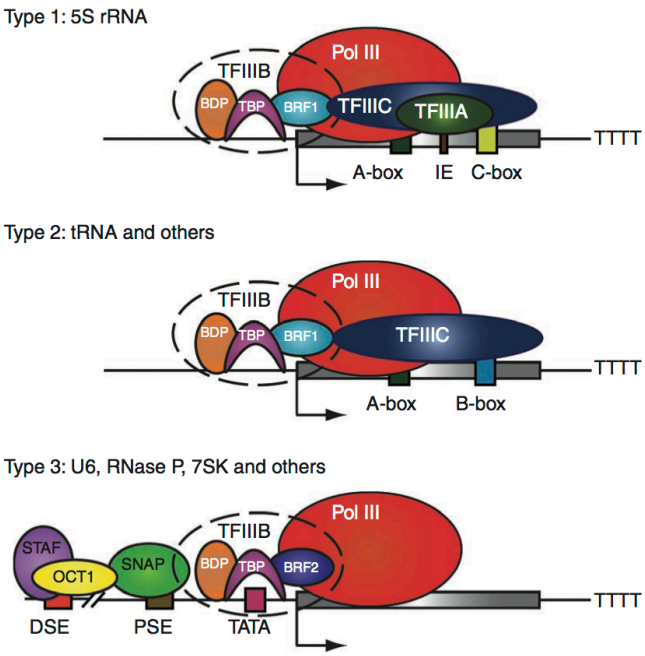


Figure 3. Pol III promoter types and associated TFs. Adapted from (18)

Recent studies however revealed that the binding of TFIIIC and TFIIIB it is not sufficient to recruit Pol III and initiate transcription at SINE loci (19). Indeed has been discovered that SINEs transcription by Pol III is selectively repressed through histone H3 trimethylation of lysine 9 (H3K9me3) by SUV39 methyltransferase which impedes Pol III recruitment whilst having much less effect on TFIIIC.
SUV39 and H3K9me3 together provide binding sites for HP1, a heterochromatin-associated protein that mediates transcriptional repression and was found at the same SINEs as SUV39H1 (Figure 4)

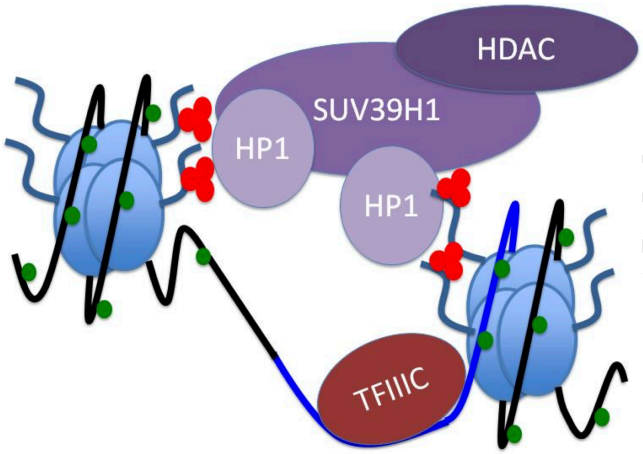


Figure 4. Model of SINE repression by SUV39H1. The blue line indicates SINE DNA while black one indicate flanking DNA. Red dots represent trimethylated H3K9. Adapted from (19)

1.4.1. *Alus*

Among SINE retrotransposons *Alus*, which are primate-specific, represent one of the most successful families, contributing almost 11% of the human genome.

Alu elements originated and began their amplification ~65 million years ago, during the radiation of primates (Figure 5). Because there is no specific mechanism for *Alu* insertion removal, new *Alu* inserts have accumulated sequence variations over time giving raise to different *Alu* subfamilies during different periods of evolutionary history. The earliest *Alu* elements where those of the J family, followed by the very active S family in which the *Alu* amplification rate reached a peak, while most of the recent *Alu* amplifications belong to the youngest Y family with Ya5 and Yb8 subfamilies being the most active in humans (20). Their current rate of retrotransposition is estimated in 1 over 20 births, which is based both on the frequency of disease-causing de novo insertions compared with nucleotide substitutions and on comparisons between the human and chimpanzee genomes and between multiple human genome sequences (14).

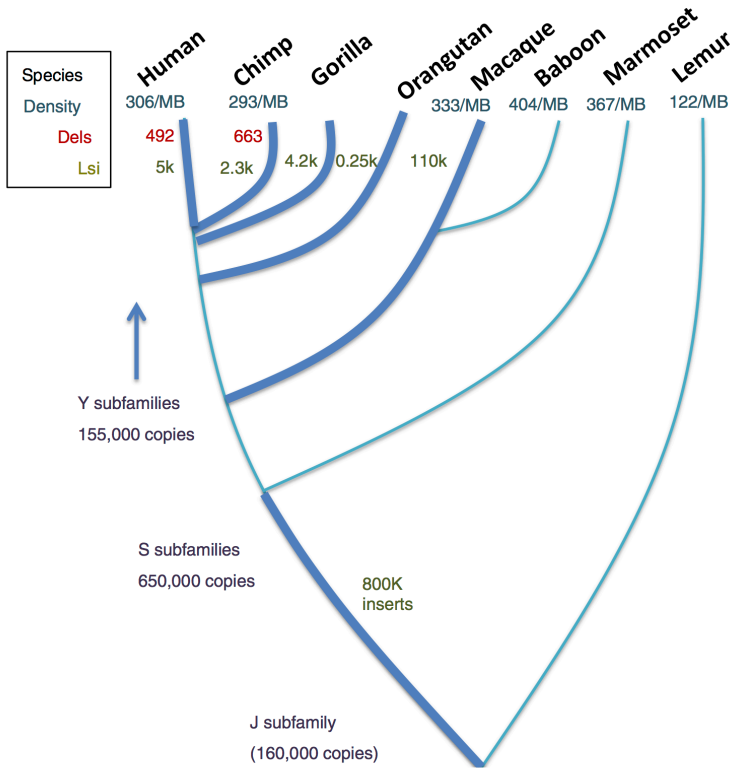


Figure 5. Evolutionary impact of *Alu* elements in primates. An approximate evolutionary tree is shown for various primate species. The approximate density of *Alu* elements is reported as the number of *Alu* elements per megabase (MB). The number of lineage-specific *Alu* insertions (Lsi) and data of *Alu*/*Alu* recombination causing deletions (Dels) between the human and chimp genomes are also shown. Adapted from (21).

The body of a typical *Alu* element is about 300 bases in length, and is formed from two diverged, 7SL-related monomers separated by a short A-rich region and is flanked by direct repeats of variable length due to the duplication of the sequences at the insertion site (21). A longer poly(A) region is located at the 3' end of the element and plays a crucial role in the retrotransposition process (22, 23). An internal, bipartite RNA polymerase (Pol) III promoter element, composed of an A and a B box both located within the left monomer, make *Alus* potential targets for the Pol III transcription machinery, which can initiate transcription at the beginning of the *Alu* and terminate at the closest poly(dT) termination sequence encountered downstream of the *Alu* body (24-26) (Fig 6).

In particular, *Alu* transcription by Pol III requires the recognition, within the *Alu* left monomer, of the internal promoter by the assembly factor TFIIIC, which in turn recruits the Pol III-interacting initiation factor TFIIIB on the ~50-bp upstream of the transcription start site (TSS) (27). Even though TFIIIB-DNA association is generally sequence-independent, an influence of the 5'-flanking region on *Alu* transcription was put in light in early studies and later confirmed in vitro and in transfected cell lines (28, 29).

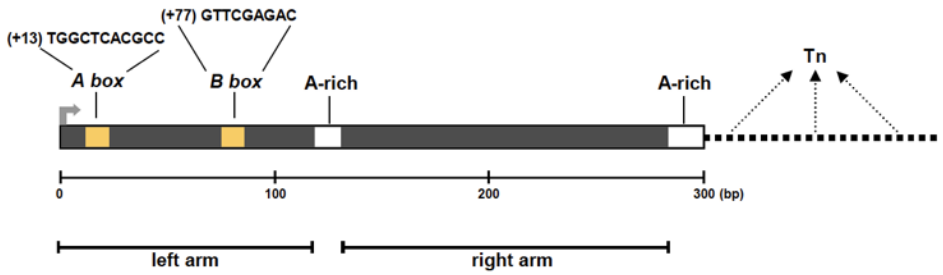


Figure 6. Architecture of *Alu* elements considered as RNA polymerase III transcription units. Schematic representation of a typical *Alu* element, approximately 300 bp in length (indicated by graduated bar). *Alu* transcription by RNA polymerase III requires A box and B box internal promoter elements (orange bars) (30), which form together the binding site for TFIIIC. The consensus sequences for *Alu* A and B boxes are reported above the scheme. While the *Alu* B box sequence perfectly matches the canonical B box sequence found in tRNA genes, the sequence of *Alu* A box slightly diverges from canonical A box sequence (TRGYnnAnnnG; (27)). Transcription is thought to start at the first *Alu* nucleotide (G) (25, 26). The A box starts at position +13, the B box 53 bp downstream, at position +77. The left and right arms of the *Alu*, each being ancestrally derived from 7SL RNA, are separated from each other by an intermediate A-rich region, starting 35 bp downstream of the B box, whose consensus sequence is A_5TACA_6. Another A-rich tract is located 3' to the right arm, at the end of the *Alu* body, starting at approximately 150 bp downstream of the middle A-rich region. Transcription termination by RNA polymerase III is expected to mainly occur at the first encountered termination signal (T*n*) downstream of the 3' terminal A-rich tract. Such a signal, either a run of at least four Ts or a T-rich non-canonical terminator (31), may be located at varying distances from the end of the *Alu* body, thus allowing for the generation of *Alu* primary transcripts carrying 3' trailers of different lengths and sequences.

Although *Alus* are repeated elements, their Pol III-synthesized transcripts are mostly unique due to accumulation of mutations in their source element, the length

and heterogeneity of the poly(A) tail and, more importantly, because of the unique 3' end transcribed from the genomic region between the poly(A) tail and the Pol III terminator. These RNAs are thought to assemble in ribonucleoprotein particles with SRP9/14 heterodimer and with poly(A) binding protein (PABP) (32, 33). These proteins are thought to help *Alu* RNAs to associate with ribosomes, likely in the nucleolus (34), with which they are exported in the cytoplasm where they compete for ORF2 protein, translated from L1 RNA, with the result of favouring its own retrotransposition to a new genomic location through the process of Target-Primed Reverse Transcription (11). While L1 depends on both its encoded proteins ORF1p and ORF2p for its mobilization, *Alu* RNA needs only ORF2p even though the presence of ORF1p enhances, either directly or indirectly, the interaction between *Alu* RNA and the factors needed for its retrotransposition (35).

1.4.2. Mammalian-wide Interspersed Repeat (MIR)

Mammalian-wide Interspersed Repeats (MIRs) represent an ancient family of tRNA-derived SINEs, found in all mammalian genomes, whose amplification seems to have ceased in the ancestors of placental mammals (36, 37).

It is thought that the MIR may have arisen following the fusion of a tRNA molecule with the 3'-end of an existing LINE (38). The complete MIR element is about 260 bp in length and possesses a tRNA-related 5' head, a 70-bp conserved central domain containing a 15-bp core sequence, two downstream segments previously described as separate interspersed repeats (MER24 and DBR (37)) and a LINE-related sequence located at the 3'-end (38) (Figure 7).

MIR elements were actively propagating prior to the radiation of mammals and before placental mammals separated. For this reason their age was originally estimated in ~130 million years (myr) even if it has been suggested that the core-SINE may have originated ~550 myr ago due to the remarkable similarity between Ter-1 (the MIR consensus in placental mammals which coincides with that revealed earlier in humans) and the OR2 SINE of octopuses (39).

Intriguingly there are observations suggesting that the core region may serve some general function in mammalian genomes, since the level of sequence conservation is higher compared to flanking 3' and 5' sequences (40).

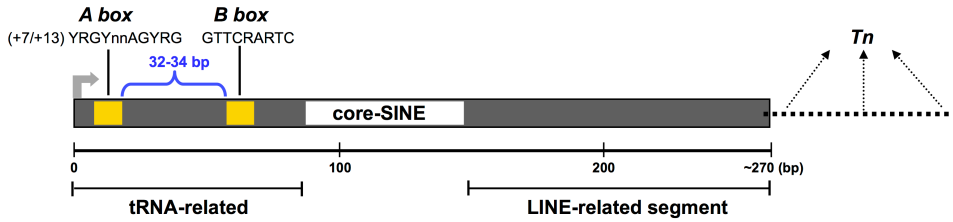


Figure 7. Representation of the structure of a mammalian-wide interspersed repeat (MIR). A tRNA-related region contains A- and B-box promoter elements driving Pol III transcription by being recognized by TFIIIC. Core-SINE indicates a highly conserved central sequence, followed by a LINE-related region. Pol III is expected to terminate at the first encountered termination signal (T*n*) which may be located at varying distances from the end of the MIR body

In the human genome there are more than 500,000 annotated MIRs (41) and have been grouped in 4 subfamilies, based on their sequence similarity, named MIR, MIRb, MIRc and MIR3. Like all SINEs, MIRs are thought to be transcribed by the RNA polymerase III machinery, with the assembly factor TFIIIC recognizing the A- and B-box internal control regions within the tRNA-derived portion of the element (37). The first experimental verifications of MIRs as Pol III targets in the human genome have come from the results of genome-wide location analysis of the Pol III machinery in human cells (18, 42) and in mouse (43, 44).

In particular, these studies revealed that, in immortalized fibroblasts, the Pol III machinery is consistently associated (with POLR3D and BRF1 TFs) with a MIR located in the first intron of the POLR3E gene, coding for a specific subunit of Pol III, and to a lesser extent (i.e. only with POLR3D) with other four MIRs (42).

Being MIRs non autonomous in retrotransposition, they could in principle exploit the LINE-encoded machinery to amplify in the genome. However, probably due to mutations in the 3' A-rich tail which prevent the binding of the endonuclease/retrotranscriptase ORF2p, they became retropositionally inactive ~130 myr ago. We can nevertheless speculate that, since part of these elements are still transcriptionally active, at least in the human and mouse genome (42) (Carnevali et al, in preparation), changes in the 3' tail due to sequence mutations could potentially rescue the retrotransposition potential of some of them.

1.5. Retrotransposition process and Target-primed Reverse Transcription (TPRT) mechanism

The mechanism through which transposons amplify in the hosting genome varies between classes.

Non-LTR retrotransposons, such as L1, *Alus,* SVA and, potentially, MIRs propagate using a mechanism analogous to target-primed reverse transcription mechanism (TPRT) established for the *Bombyx mori* R2 element (45) which involves the two proteins ORF1p and ORF2p encoded by L1 elements.

During the initial step of the TPRT process (Figure 8), a functional L1 element is transcribed by Pol II from its internal promoter and its bicistronic mRNA is brought into the cytoplasm where the two aforementioned ORFs (ORF1 and ORF2) are translated. The two L1 encoded proteins, which show a *cis* preference for the RNA encoding them (8, 46), bind the L1 RNA molecule forming a ribonucleoprotein (RNP) complex, which subsequently travels to the nucleus where the insertion occurs.

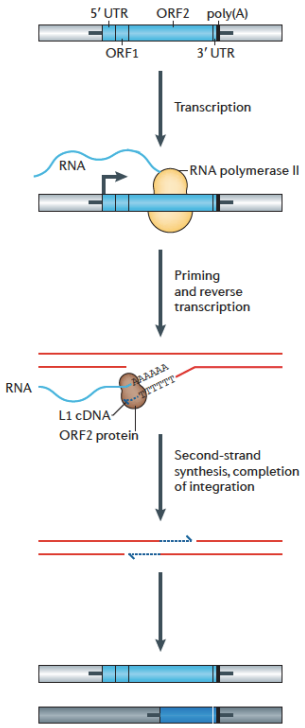


Figure 8. Target-primed reverse transcription mechanism. Adapted from (47)

This *cis* preference may decrease the ability of other mRNA to access ORF1p and ORF2p explaining the lower retropseudogene copy numbers in the genome. However *Alu* RNAs that bind to ribosomes thanks to SRP9/14, can compete more efficiently with L1 for ORF2p and thus can easily retrotranspose, also given their lack of ORF1 requirement (48).
It is possible that ORF1p, which is present in higher numbers than ORF2p, helps displacing L1 RNA from the ribosome and preventing it from entering the RNA degradation pathways thus promoting entrance in the nucleus through a mechanism still to be elucidated (49).
On the other hand another study discovered that in the cytoplasm of mammalian cells, ORF1p and non-LTR retrotransposons RNPs are localized to stress granules, cytoplasmic bodies closely associated with P-bodies, suggesting a mechanism by which the cell may mitigate the mutagenic effects of L1 retrotransposition by sequestering L1 RNPs and possibly targeting them for degradation (50).

Once the RNP arrives in the nucleus, ORF2p with its endonuclease activity cuts the genomic DNA (gDNA) leaving a 3'-OH terminus. The cut occurs with low sequence specificity, but (dT_n-dA_n) sites are preferentially recognized. The 3' end of the RNA retrotransposon, which contains the poly(A) tract, anneals to the dT_n nicked gDNA where the 3' exposed hydroxyl is then used as a primer for first strand synthesis by ORF2-encoded reverse transcriptase (8). A nick, staggered to the first one, occurs on the second strand by unknown mechanism and second-strand synthesis is primed. The target site is now filled with the cDNA copy of the original retroelement and the single-strand DNA (ssDNA) remaining at the target sites is filled producing target site duplications (TSDs) (Figure 8).

1.6. Retrotransposons and human genome evolution

TEs are probably the most powerful genetic force that drove evolution in higher species. Among these, non-LTR retrotransposons (LINEs, *Alus*, SVA) contributed most to human genome evolution due to their continuous activity over tens of

millions of years that led to an accumulation of these elements concurrently with increased genome size (14).
The processes mediated by TEs are known to be sources of local genomic instability, structural variations and genomic rearrangements as well as genetic innovation and gene expression regulation events.

The most straightforward way a retrotransposon can alter genome function is by inserting into a protein coding gene or regulatory regions having either deleterious or beneficial effects on his host. Thus retrotransposons can perturb gene expression by disrupting exons, introducing alternative splicing signals (a process termed exonization), as well as polyadenilation signals. Moreover, being the retroelement sequence a 'sink' of transcription factor binding sites, retrotransposon insertions has also been exapted by the nearby genes thus starting to act as enhancers, insulators or promoters (reviewed in (51)). In recent studies indeed *Alus* and MIRs (which are currently inactive in their retrotransposition process) have been found to be enriched in chromatin regions presenting marks typical of enhancers (41, 52) although nothing is known about the possible involvement of *Alu* and MIR expression in enhancer function.
Another study found that numerous MIR sequences serve as insulator across the human genome, serving as both chromatin barrier activity and enhancer-blocking activity. The first role appears to be cell-type specific while the second appears to be conserved across cell types and between species (53).

In addition to insertions, retrotransposons can lead to structural variations and genomic rearrangements through the mechanism of non-allelic homologous recombination that can cause deletions, segmental duplications and inversion of the involved genomic regions (reviewed in (14)).
To date 96 retrotransposition events (25 L1, 60 *Alu*, 7 SVA, or 4 poly(A)) are known resulting in single-gene disease (reviewed in (54)). Among them, one of the very first to be discovered was the insertion of an L1 element in the *FVIII* gene causing Hemophilia A (55).

Alus are often located into or nearby genes (either coding or non-coding) and their sequence can be target of other regulators that affect the expression of the hosting

gene, both at transcriptional and post-transcriptional level. Thus they are capable of epigenetic perturbation, being their genomic sequence target of noncoding RNA which induce the deposition of repressive epigenetic modifications (56), are implicated in mRNA decay mediated by Staufen1 (57) and mediate gene silencing through the ADAR editing process (58).

Retrotransposons are also capable of releasing regulatory RNAs such as micro RNAs (miRNAs), small interfering RNAs (siRNAs) and piwi-interacting RNAs (piRNA) (59-62), which interestingly can also act as an additional host mechanism through which it controls their activity.

1.6.1. Host mechanisms controlling retrotransposons

Even if it is now accepted the beneficial role that TEs played in human genome evolution, an 'out-of-control' spread of these elements would rapidly lead to lethality. Thus higher species have evolved different mechanism to tame their amplification achieving a fine balance between potentially deleterious events and adaptive benefits of genetic diversity (Figure 9). Host control mechanisms have been reported for all types of transposable element. In this paragraph, discussion will be limited to retrotransposons.

One of the most documented mechanism through which the cell domesticated retrotransposons is by depositing locus-specific repressive chromatin marks, which are enriched at TE loci, mostly through DNA methylation which has been proposed to be evolved as defense mechanism against transposable elements. In addition to DNA methylation, the host represses retrotransposons activity by the methylation of histone H3 lysine 9 at their loci (reviewed in (5, 63)) (64). H3K9 methylation (and not DNA methylation) has recently been proposed to be the primary mechanism for SINE transcription repression (65).

As mentioned above, cells control the amplification of retrotransposons also through the RNA-induced silencing mechanism. Intriguingly small noncoding RNAs (in

particular piRNAs and siRNAs) that are key players in this mechanism can originate from the retrotransposons transcripts themselves (61, 66, 67).

Another 'weapon' of the cell's arsenal to control retrotransposons activity is nucleic acid editing process. APOBEC3A/B proteins, which are members of the APOBEC3 family of human innate antiretroviral resistance factors, can enter the nucleus, where LINE-1 and *Alu* reverse transcription occurs, and specifically inhibit both LINE-1 and *Alu* retrotransposition. Although the factors involved are cytidine deaminase, the mechanism through which they inhibit L1 and *Alu* retrotransposition still has to be elucidated (68).

As a further element of connection between retrotransposon control and RNA-mediated silencing, Microprocessor (Drosha-DGCR8), a nuclear protein complex involved in miRNA biogenesis, has been shown recently to control the activity of retrotransposons by binding L1, *Alu*, and SVA-derived RNAs in human cells (60).

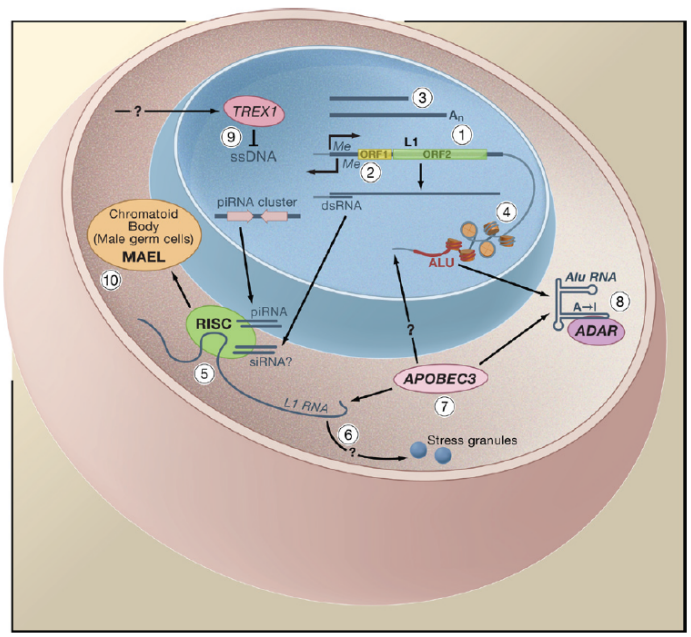


Figure 9. How the cell affects retrotransposons. Shown are the various mechanisms used by the cell to domesticate TEs. See text for description, Adapted from (5).

1.7. *Alu* and MIR expression in different cell types and conditions

Despite the abundance of SINE elements in the human genome, and the generally high efficiency with which Pol III transcribes its target genes (69), the cellular levels of Pol III-synthesized transcripts of *Alus* and MIRs are usually very low in normal conditions and, accordingly, most of them are not occupied by the Pol III transcription machinery in human cells (70). *Alu* and MIR transcription by Pol III is thought to be limited by epigenetic silencing mainly involving histone methylation (65), but a satisfactory picture of their expression regulation at the transcriptional level is still lacking (71). In particular, early and recent studies have shown that *Alu* transcription by Pol III may be deregulated in response to different types of signal but the involved molecular mechanisms are still largely unknown. To date no studies on MIR expression regulation have been carried on probably due to the fact that these elements lost their retrotransposition activity ~130 million years ago.

One of the most documented signals that strongly stimulate transcription of endogenous *Alu* elements transcribed by Pol III in humans is viral infection by herpes simplex virus type 1 (HSV-1) and Adenovirus 5 (Ad5) (72, 73). In these studies, it was suggested that virus-dependent induction of *Alu* expression may be mediated through *Alu* internal regulatory sequences. According to a proposed mechanistic hypothesis, viral components can modulate the activity of factors interacting with the core intragenic type II promoter present in the majority of *Alu* elements.

The abundance of Pol III transcribed *Alu* RNAs is also known to be transiently increased during heat shock and cycloheximide stress stimuli indicating that induction of *Alu* expression is a general cell response to stress (74). In response to heat shock indeed *Alu* RNAs have been shown to bind RNA polymerase II (Pol II) and repress transcription of a subset of genes. *Alu* RNA prevent Pol II from properly engaging the DNA during closed complex formation, resulting in complexes with an altered conformation that are transcriptionally inert (75).

Interestingly, the effects on *Alu* transcriptome of adenovirus infection, heat shock and cycloheximide, seem to be both cell-type and condition-dependent. Indeed researchers found that in K562 cells (immortalized cell line produced from a female patient with chronic myelogenous leukemia (CML)), which are unusually hypomethylated and in which *Alu* repeats are far more actively transcribed than those in other human cell lines and somatic tissues, the level of *Alu* RNAs were relatively insensitive to these stress stimuli. In the same study it was also found that transcription of transiently transfected *Alu* templates was repressed by methylation in all cell lines tested but cell stresses were not able to relieve this repression suggesting that they activated *Alu* transcription through another, still unexplored pathway (76).

Alu RNAs have also been found to be involved in the regulation of stem-cell proliferation (62). Recent bioinformatics analyses discovered one subset of *Alu* repeats that harbors the characteristic 6-bp core retinoic acid receptor (RAR) binding site (direct repeat) spaced by two nucleotides called the DR2 element (77). The ~2-3% of the ~100,000-200,000 DR2 element-containing *Alu* repeats located close to activated Pol II genes are activated by RAR in human embryonic stem cells to generate Pol III-dependent RNAs. These transcripts are further processed into small RNAs (~28-65 nt) that target a subset of crucial stem-cell mRNAs causing their degradation and modulating exit from the proliferative stem-cell state. This phenomenon has been discovered in Ntera2 cells but not in other cell types such as HeLa cells or human lung fibroblasts, suggesting a cell type/condition-dependent mechanism

The *Alu* transcriptome has also been found to be deregulated in response to growth factors. CGGBP1 is a repeat-binding transcription regulatory protein that regulates, among others, cell proliferation and growth as well as stress response. It has been discovered that it binds *Alu* elements impeding RNA Pol III binding and suppressing *Alu* transcription in *cis*. CGGBP1 depletion, following serum stimulation, increases *Alu* RNA levels and also negatively impact, in a way that mimics heat shock in terms of gene expression changes, the amount of protein-coding mRNA through *Alu*-mediated inhibition of RNA Pol II activity. Thus

CGGBP1 affects global gene expression through regulation in cis of *Alu* RNA levels, which in turn affects RNA Pol II in *trans* (78).

1.8. SINE expression profiling

Studies on *Alu* and MIR elements expression profiling has never been carried on in depth, because the identification of genuine Pol III-transcribed *Alu* and MIR RNAs is hampered by two main problems: (i) the extremely high copy number and sequence similarity of *Alu* and MIR elements within the human genome, and (ii) their frequent location inside introns or untranslated regions of primary or mature Pol II transcripts.

To date, while no studies have been conducted on MIR expression regulation, those aimed at identifying Pol III-derived *Alu* RNAs have been performed using low throughput techniques such as Northern hybridization, allowing to distinguish them from *Alu* RNA passenger of longer Pol II transcripts, and C-RACE followed by sequencing of the unique 3' ends to identify source loci of transcription. Even if Northern hybridization is effective in global *Alu* RNA quantification, it fails in assessing the expression level of individual *Alu* loci, while techniques used to identify transcriptionally active *Alu* elements are unfeasible on a genomic high throughput scale.
Recently the development of Next-generation Sequencing (NGS) techniques has been exploited, through the use of Chromatin Immuno-Precipitation followed by massive parallel sequencing (ChIP-seq), to identify transcriptionally active *Alu* loci, whose association with components of the Pol III machinery has been used as an evidence of transcription. However, even if this high-throughput technique has the advantage to be carried out on a genomic scale, the association of Pol III transcription factors (TFs) to an *Alu* element does not necessarily indicate its transcription. Quantification of expression levels is also not feasible through this approach. Therefore, none of the above mentioned approaches could be used for a comprehensive and quantitative expression profiling of SINEs.

A recent NGS application, called RNA-seq, has been developed for transcriptome profiling. This approach provides a far more precise measurement of transcripts levels and their isoforms than other methods (e.g. Microarray) and is able to identify new splice variant as well as new non-coding transcripts (reviewed in (79, 80)).

However when RNA-seq, as well as other NGS techniques, has to face with repetitive elements, an important problem arises related to read mapping at these genomic loci. Indeed NGS technology producing high data volume of relatively short sequencing reads (~50-150 bp in length) have made this challenge more difficult. Repeats, from a computational perspective, create ambiguities in their alignment and assembly, leading to biased results. Nevertheless solutions have been proposed to solve this issue with NGS technology, while other methods such as Microarrays cannot deal with it being impossible for them to map transcripts arising from repeat elements.

To address this problem three strategies for read mapping have been proposed: i) "unique", which reports only reads mapping uniquely on the genome; ii) "best match", which reports the best possible alignment for each read, determined by the scoring function, and, in case of equally mapping scores, it reports one randomly; iii) "all matches", which reports all possible alignment, including the low-scoring ones (81).

The choice of one strategy versus the other depends on the goals of the experiment and on the type of sequencing reads available. Indeed NGS technologies are evolving producing longer and paired-end reads (i.e. reads coming from the sequencing of both ends of the same cDNA fragment, which in turns consists in a virtual longer sequencing coverage) as well as reads maintaining the strand information of the transcript from which they arise.

Thus, given an RNA-seq dataset of sufficiently long reads (e.g. 75 nt-long paired-end reads), the "unique" strategy seems to be the most appropriate for SINE expression profiling. Indeed most of the *Alu* and MIR elements contain at least very few sequence variations throughout their entire length that make them almost unique. Thus these small differences in sequence could be exploited by sufficiently long RNA-seq reads such as those discarded due to multi-mapping would be few

and would not affect their identification while affecting only partially their quantification.

However a combined strategy of “unique” and “best match” alignment strategy could be used to search for rare identically in sequence transcriptionally active SINE elements.

2. Goals of this study

To overcome the limitations of previous strategies for the study of SINE expression, we developed a bioinformatic pipeline that exploits RNA-seq data to reveal genuine Pol III-transcribed SINE loci.
Our studies were thus aimed at profiling for the first time *Alu* and MIR expression in different human cell types and conditions, in order to reveal changes in SINE expression profiles correlating with different cellular states. Such profiles in turns could be used, along with data from other types of epigenomic profiling, such as ChIP-seq against TFs and histone modifications as well as Pol II gene expression profiling, to leverage our understanding of the mechanisms behind their regulation as well as their role in cell specificity and pathological conditions.

In particular we tried to exploit our bioinformatics pipeline for the study of:

- *Alus* and MIRs expression profiling in seven ENCODE cell lines
- *Alus* expression dysregulation in a model of viral oncogenesis (IMR90 *dl*1500 Ad5 infected cells)
- *Alus* expression dysregulation in human cancer (gastric adenocarcinoma from The Cancer Genome Atlas)

3. Materials and Methods

3.1. Datasets

The annotated *Alu* and MIR elements considered in our studies where downloaded from the Repeatmasker track in the UCSC Table browser for the human genome version GRCh37/hg19. Listed below are the main NGS datasets used in our studies for expressed SINE loci identification.

3.1.1. ENCODE

The Cold Spring Harbor Lab (CSHL) long RNA-seq data within ENCODE (whole-cell polyA+ and polyA- RNAs, two replicates for each sample) relative to the following cell lines: Gm12878, H1-hESC, HeLa-S3, HepG2, HUVEC, K562, NHEK, for a total of 28 datasets, were used for *Alu* and MIR RNAs identification, (http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeCshlLongRnaSeq). These datasets contain stranded paired-end reads (2x76 nucleotides long). We also made use of ChIP-seq peak data for some Pol III TFs from ENCODE/Stanford/Yale/USC/Harvard (http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeSydhTfbs/) as well as ChIP-seq peak data for a plethora of other TFs from (http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeRegTfbsClustered/ wgEncodeRegTfbsClusteredWithCellsV3.bed.gz)

3.1.2. *dl*1500 Ad5 infected IMR90 cells

For the identification of genuine Pol III *Alu* RNAs we made use of 100 nt-long paired-end stranded reads generated by Illumina HiSeq 2000 sequencing of total RNA extracted from infected fetal lung fibroblasts. Contact-inhibited G1-arrested fetal lung fibroblast primary cells (IMR90, not immortalized) were harvested at 6 and 24 hours post infection with *dl*1500, a modified version of Adenovirus 5 (Ad5)

expressing only the small e1a protein with little or no expression of other viral genes (82). Control cells were only grown in media.
We also used data from ChIP-seq performed against POLIII (RPC39), BDP1 and TFIIIC-110 factors in IMR90 cells 24 hours post-infection (unpublished data collected in collaboration with Arnold J. Berk laboratory, Department of Microbiology, Immunology, and Molecular Genetics, Jonsson Comprehensive Cancer Center, UCLA) as well as ChIP-seq against various histone modifications including acetylation of histone H3 lysine 27 (H3K27ac) (83), lysine 9 (H3K9ac) and lysine 18 (H3K18ac) from previous studies on the effects of Adenovirus small e1a oncoprotein on the reorganization of the host epigenome (84), and P300 (lysine acetylase) and RB1 (retinoblastoma) proteins (83).

3.1.3. TCGA (The Cancer Genome Atlas)

We used polyA+ Illumina HiSeq 2000 RNA-Seq raw data (fastq) consisting in 2x75 nucleotides long unstranded paired-end reads, of 72 samples belonging to 36 different patients, each with 1 tumor and 1 non tumor sample, affected by Stomach adenocarcinoma (STAD) (85). Raw data were downloaded from the Cancer Genomics Hub (https://cghub.ucsc.edu/).

3.2. Bioinformatic pipelines for individually expressed SINE identification

During our research studies we designed a first version of a bioinformatic pipeline, to identify individually expressed SINE transcripts consisting in a *shell script* aimed at automating the informatics data flow across several publicly available (open source) tools.
Afterwards we improved the pipeline developing a Python script that let us more control on data manipulation and filtering (from here on called "*SINEsFind*").
The first bioinformatic pipeline was used to identify genuine Pol III-derived *Alu* transcripts in ENCODE datasets while the improved one was used for MIR

expression profiling in the ENCODE datasets as well as *Alu* expression profiling in the *dl*1500 Ad5 IMR90 infected cells and TCGA datasets (see below).
An outline of these two pipelines is provided in this section. A more detailed description of computational methods can be found in Supplementary Materials.

3.2.1. First developed bioinformatic pipeline

Reads from each dataset were aligned to the reference genome (GRCh37/hg19) using TopHat aligner (86) with default settings (allowing to retain reads with up to 20 equally scoring hits in the genome). Uniquely aligned paired-end reads (identified by NH:i:1 in the alignment file) were recognized and counted, for each annotated *Alu*, through the htseq-count tool of the HTSeq Python package (87). Only *Alus* with a number of mapped reads over the calculated background noise were retained (see Supplementary Materials). To check for the performance of the aligner and its reliability in unique alignment, we replaced TopHat by the independently developed STAR aligner (88) for the analysis of two datasets (NHEK polyA+, replicates 1 and 2), and found largely (~96%) overlapping sets of *Alus* with more than 10 uniquely mapped paired-end reads.
The coordinates of retained *Alus* were supplied to sitepro script of the Cis-regulatory Element Annotation System (CEAS suite; http://liulab.dfci.harvard.edu/CEAS/) along with the corresponding RNA-seq stranded signal profiles. We used sitepro (developed mainly for ChIP-seq data) because it allowed us to calculate the signal profile in a range of +/- 500 nt from the center of the *Alu* body with a resolution of 50 nt. In this way we could address the problem of 'passenger' *Alu* RNAs by devising a filter aimed at excluding false positives on the basis of the level of upstream and downstream spurious RNA signals (see Supplementary Materials for details). Figure 10 shows a schematic representation of the pipeline.

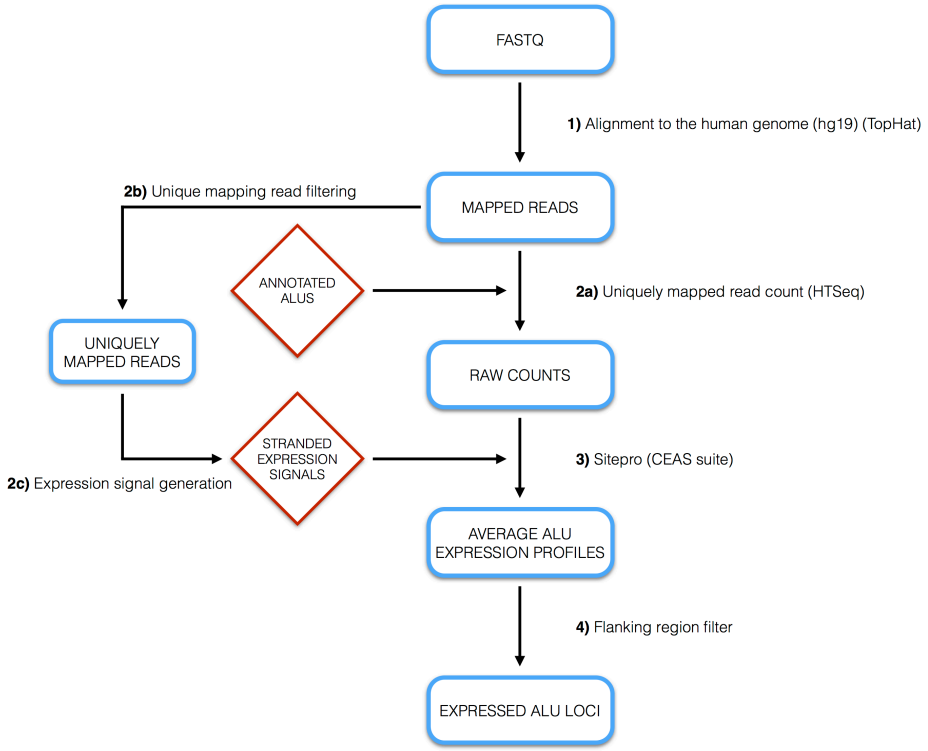


Figure 10. *Alu* **RNA identification pipeline.** Shown is a flow-diagram of the first developed bioinformatic pipeline for the identification of autonomously expressed *Alu* loci from RNA-seq data sets. See Results and Materials and Methods for details.

3.2.2. Improved bioinformatic pipeline: *SINEsFind*

The bam files from each dataset containing RNA-Seq reads aligned to the reference genome (GRCh37/hg19) using TopHat aligner, along with the annotated SINEs of interest, are submitted to the 'in house' developed *SINEsFind* Python script to perform the identification of individual SINE transcripts.

The Python script first builds stranded coverage vectors for the whole genome, using the bam file supplied using uniquely mapped reads (tag NH:i:1 in the bam file). Then, for each annotated SINE having an expression coverage value over a calculated background noise threshold, the script calculates the coordinates of the corresponding expected full-length consensus element (see Supplementary Materials), to take into account the fact that many of the annotated SINE elements

are truncated,. Finally a filter (flanking region filter) is applied to the identified expected full-length element, as described in Supplementary Materials, in order to exclude false positive arising from SINE elements embedded in Pol II transcripts. Basically the filter aims to do this by imposing a significant lower expression coverage value to the flanking regions immediately upstream and downstream of the expected full-length SINE, thus discriminating between genuine SINE RNAs and those 'passenger' of longer Pol II transcripts or part of their trailers extending downstream their annotated 3'UTRs. Figure 11 shows a schematic representation of this improved pipeline.

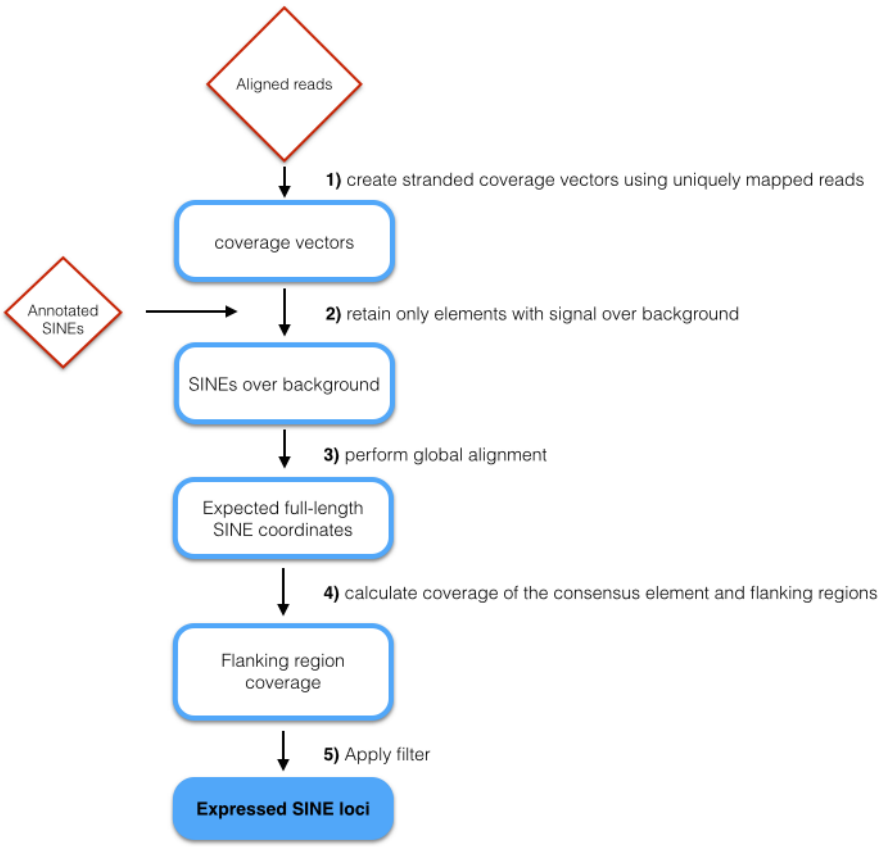


Figure 11. *SINEsFind* bioinformatic pipeline flowchart. Shown is a flow-diagram of the improved bioinformatic pipeline for the identification of autonomously expressed SINE loci from RNA-seq data sets.See Results and Materials and Methods for details.

3.3. ENCODE *Alus*: methodological add-ons

To identify genuine *Alu* transcripts in the ENCODE dataset the first developed bioinformatic pipeline (par. 3.2.1) was used along with all the annotated *Alus* (Figure 10). Only *Alus* with more than 10 mapped reads that passed the final filter of the pipeline in both ENCODE RNA-seq replicates were considered to represent autonomously expressed *Alu* loci (as such, they will be often referred to in the text as "expression-positive"). Complete lists of these *Alus* are reported in Supplementary Table S1. The bam files containing the alignments with uniquely mapped (NH:i:1) paired-end reads, generated through TopHat for all the 28 ENCODE datasets, and through STAR for a subset of them (NHEK polyA+ replicates 1 and 2; HeLa-S3 polyA+ replicate 1; K562 polyA- replicate 1), are deposited at the following link: http://bioinfo.cce.unipr.it/NAR-02564-Z-2014/. Also available at the same link is the above described pipeline in the form of a collection of shell scripts designed to automate the execution of the different publicly available software (such as TopHat and htseq-count, as detailed in the Supplementary Materials along with their specific options).

As a number of *Alu* transcripts were found both in the polyA+ and polyA- datasets, Supplementary Table S1 also contains a non-redundant list of all expressed *Alus* obtained by merging expression-positive *Alus* found in the polyA+ and polyA- fractions of all cell lines ("All non-redundant" sheet in Supplementary Table S1).

All analyses were carried out using GRCh37/hg19 genome assembly. Even though the contribution of novel sequence in GRCh38 assembly, that is absent from GRCh37/hg19, to *Alu* expression profiles was expected to be limited (the total number of bases in GRCh37 being increased by ~2% only with respect to GRCh37/hg19), we nevertheless screened a pair of ENCODE RNA seq dataset replicates (NHEK polyA+, r1 and r2) with our pipeline using GRCh37 assembly as a reference for read mapping, and compared the results with those obtained with hg19 genome assembly. We found that the vast majority (92-95 %) of *Alus* detected as expression-positive in either genome assembly was shared with the other one.

3.3.1. Additional ChIP-Seq data analyses

To further support the identification of unique *Alu* transcripts found in Hela-S3 and K562 cells, we intersected the ChIP-seq peaks of the Pol III machinery components TFIIIC-110, POLIII (RPC155 subunit), BRF1, BRF2, BDP1, derived from ENCODE/Stanford/Yale/USC/Harvard ChIP-seq data (http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeSydhTfbs/) with the expression-positive *Alu* coordinates, extended to 200 bp upstream, of these two cell lines. P-values for the association of each Pol III component to expression-positive intergenic *Alus* were calculated using the Fisher's exact test against total (intergenic) *Alus*. The lists of Pol III-associated, expression-positive *Alus* are reported in Supplementary Table S2.

To identify other transcription factors (TF) associated to expression-positive *Alu* elements, we intersected, for each cell line, the 500 bp upstream of the *Alus* with the coordinates of the TF binding sites from ENCODE ChIP-seq (http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeRegTfbsClustered/wgEncodeRegTfbsClusteredWithCellsV3.bed.gz). P-values for the association of TFs to expression-positive *Alus* were calculated using the Fisher's exact test against total *Alus*. Lists of TF-*Alu* interactions are reported in Supplementary Table S3.

3.4. ENCODE MIRs: methodologically add-ons

For MIR RNAs identification in the ENCODE dataset we used the improved version of the bioinformatic pipeline *SINEsFind* (par. 3.2.2). We decided to merge PolyA+ and PolyA- datasets (*bam* files) to streamline the analyses since we were not interested in the differences between these two features because MIR RNAs usually do not have a poly(A) tail.

Only MIRs with a coverage peak value greater than 10 where considered and among these, only those which passed the final filter of the *SINEsFind* Python script in both ENCODE RNA-seq replicates where considered to represent autonomously expressed MIR loci. Complete list of expression-positive MIRs are reported in Supplementary Table S5.

To further support the identification of unique MIR transcripts found in Hela-S3 and K562 cells and to identify other transcription factors (TFs) associated to expression-positive MIR elements, we followed the same procedures as for ENCODE *Alus* (par. 3.3.1). However, due to the use of improved bioinformatic pipeline, the new coordinates used for intersection with Pol III TFs and other coordinates, where calculated with respect to the expected full-length MIR elements (see Supplementary Materials).
The lists of Pol III-associated expression-positive MIRs and those associated to other TFs are reported in Supplementary Table S6 and S7.

Following the recent observation that MIRs are highly concentrated in enhancer of K562 and HeLa human cell types (41) we investigated if expression positive MIRs were more or less enriched in this state than all the other annotated MIRs.
We intersected our expression positive MIRs in K562 and HeLa-S3 cell lines with the coordinates of the reported MIR-enhancers (lifted over to GRCh37/hg19) and we did not find any overlap.
We also tested for enrichment in enhancer states (weak and strong enhancers) versus other states of the chromatin of the annotated MIR used in the bioinformatics pipeline, using Chromatin State Segmentation by Hidden Markov Model (HMM) from ENCODE for each cell line (excluding HeLa-S3 for which data are not available) but we found no enrichment in enhancers states of the chromatin (see Supplementary Materials).

Again we tested for the enrichment of the expression-positive MIRs in the enhancer states of the chromatin performing a Fisher's exact test against all the other annotated MIRs used in the pipeline. We found enrichment only for Gm12878 and HepG2 cells in 'strong enhancer' state (Pval $1.876e^{-05}$ and Pval $2.133e^{-05}$ respectively) (Supplementary Table S8) (see Supplementary Materials).

To investigate whether expression of MIR elements located inside Pol II genes in antisense orientation was correlated with the expression of the hosting gene, we calculated normalized reads count for each of these MIRs and the corresponding hosting genes using htseq-count and DESeq2. We then calculated expression correlation using Pearson coefficient (see Supplementary Materials).

3.5. *dl*1500 Ad5 infected IMR90 cells

To identify genuine *Alu* transcripts in dl1500-infected IMR90 cells, we applied to each dataset the *SINEsFind* bioinformatic pipeline (par. 3.2.2) limiting the analysis to *Alus* annotated in intergenic regions and inside Pol II genes but in antisense orientation (from here on named "intergenic/antisense"). Only *Alu* elements with a peak value of expression coverage greater than 5 (see Supplementary Materials) were further processed in the script with the Flanking Region Filter and only those which passed the filter where retained and considered as *Alus* likely to be expressed as autonomous transcription units. The complete list of these *Alus* is reported in Supplementary Table S9.

To support their genuine Pol III transcription, the coordinates of the corresponding full-length *Alus* calculated using *SINEsFind* (see Supplementary Materials) were extended to 200 bp upstream and intersected with the coordinates of ChIP-seq peaks of the Pol III machinery component RPC39, TFIIIC-110 and BDP1. The list of Pol III-associated expression-positive *Alus* is reported in Supplementary Table S10.

Moreover we tested for enrichment in H3k9ac, H3k18ac and H3k27ac of the expression-positive *Alu* elements again intersecting the coordinates of the corresponding expected full-length *Alus* with those of the histone modifications ChIP-seq peaks. To test the enrichment of expression-positive *Alus* in P300 and RB1 proteins we intersected the coordinates of the 500 bp upstream the expected full-length *Alus* with those of protein peaks.

All the P-values for the association of each of these factors to expression-positive *Alus* were calculated performing Fisher's exact test against the whole dataset of annotated *Alus* used.

3.6. TCGA

Expression-positive *Alus*, with an expression coverage peak value greater than 10 (the calculated average background noise value, see Suplementary Materials), resulting from the application of the improved *SINEsFind* bioinformatic pipeline

(par. 3.2.2.) to all the 72 datasets (referred to in par. 3.1.3) are listed in Supplementary Table S11.
For each patient all the tumor and non-tumor *Alus* were also merged in a single non-redundant list and the expression coverage area of each corresponding expected full-length *Alu* was calculated both for the tumor and non tumor samples and normalized by Total Count method (see Supplementary Materials). The calculated coverage areas of each *Alu* were summed to roughly quantify the genuine (i.e. not due to longer host transcripts) *Alu* expression in tumor and non-tumor samples.

3.7. DNA constructs and in vitro transcription

3.7.1. *Alu* plasmid construction

Using oligonucleotides listed in Supplementary Table S4, nine human *Alu* loci (whose chromosome coordinates are reported in Table 1), together with 5'- and 3'–flanking regions, were PCR-amplified from buccal cell genomic DNA with GoTaq® DNA polymerase (Promega) and cloned into pGEM®-T Easy vector (Promega). Constructs containing targeted mutation of the B box internal control element were obtained by recombinant PCR through the fusion of sub-fragments overlapping in the mutated region, as previously described (89), followed by cloning into pGEM®-T Easy. Upstream deletion constructs employed forward PCR primers generating amplicons truncated to position -12 (or -15, in the case of *Alu*Sx_chr10) with respect to *Alu* 5' end. Truncated amplicons were inserted into pGEM®-T Easy; the constructs selected for *in vitro* transcription contained the 5'-truncated insert with the same orientation as its wild type *Alu* counterpart, to minimize the influence of vector sequence on transcription efficiency.

3.7.2. MIRs plasmid construction

Four human MIR loci (whose chromosome coordinates are reported in Table 2), together with 5'- and 3'-flanking regions, were PCR-amplified and cloned into pGEM®-T Easy vector (Promega) using oligonucleotides listed in Supplementary Table S13, as described above for *Alus*. Constructs containing targeted mutation of

the B box internal control elementwere also obtained as described above for *Alus*. Upstream deletion constructs employed forward PCR primers generating amplicons truncated to position -25 with respect to MIR A box at the 5'-end. Truncated amplicons were inserted into pGEM®-T Easy and the constructs selected for *in vitro* transcription, contained the 5'-truncated insert, had the same orientation as its wild-type MIR counterpart, to minimize the influence of vector sequence on transcription efficiency.

3.7.3. *Alu* and MIR in vitro transcription

All recombinant plasmids for *in vitro* transcription reactions were purified with the Qiagen Plasmid Mini kit (Qiagen). Reaction mixtures (final volume: 25 µl) contained 500 ng of template DNA, 70 mM KCl, 5 mM MgCl2, 1 mM DTT, 2.5% glycerol, 20 mM Tris–HCl pH 8, 5 mM phosphocreatine, 2 µg/ml alpha-amanitin, 0.4 U/µl SUPERase-In (Ambion), 40 µg of HeLa cell nuclear extract(90), 0.5 mM ATP, CTP and GTP, 0.025 mM UTP and 5µCi of [α-32P]UTP (Perkin- Elmer). Reactions were allowed to proceed for 60 min at 30°C before being stopped by addition of 75 µl of nuclease free water and 100 µl of phenol:chloroform pH 5.5 (1:1). Purified labeled RNA products were resolved on a 6% polyacrylamide, 7 M urea gel and visualized and quantified with the Cyclone Phosphor Imager (PerkinElmer) and the Quantity One software (Bio-Rad).

Table 1. *Alus* subjected to in vitro transcription analysis

Alu	Expression in cell lines[1]	Predicted length of primary transcript(s)[2]
AluSq2_chr1 (chr1:61523296-61523586)	H1-hESC, HeLa-S3, Hep G2, K562, NHEK	**355** (T_4); **361** (T_{10}).
AluSx_chr1 (chr1:235531222-235531520)	none	**328** (TAT_3); **338** (TAT_3); **431** (T_4)
AluSx1_chr3 (chr3:139109300-139109588)	H1-hESC, GM12878 (sporadical)	**304** (T_3GT); **311** (TCT_3); **437** (TAT_3); **443** (T_{17})
AluY_chr7 (chr7:73761603-73761897)	K562 (sporadical)	**322** (T_5)
AluY_chr10-a (chr10:103929441-103929803)	H1-hESC (sporadical)	**370** (TCT_3); **376** (T_4); **397** (T_6); **406** (T_3GT_2)
AluY_chr10-b (chr10:69524852-69525156)	NHEK	**397** (T_5)
AluSx_chr10 (chr10:12236879-12237173)	none	**320** (T_4); **456** (T_6)
AluSp_chr17 (chr17:4295121-4295437)	K562	**387** (T_3CT); **424** (TAT_3); **430** (T_6)
AluY_chr22 (chr22:41932115-41932411)	none	**378** (TGT_3); **409** (T_4); **590** (T_3CT)

[1] This column lists, for each *Alu* element, the cell lines in which it was found to be expressed by RNA-seq data analysis.

[2] The reported transcript lengths were calculated by assuming as TSS the G at the first *Alu* position, located 12 bp upstream of the T with which the A box starts (TRGY...). This assumption is based on early *in vitro* transcription analyses showing that most *Alu* transcripts initiate in close proximity to the 5' end of the consensus *Alu* sequence. To estimate the 3' end of the transcript, both canonical (Tn with n≥4) and non-canonical T-rich Pol III terminators were considered downstream of *Alu* body sequence (indicated in parentheses after the transcript length); for canonical terminators, the 4 Us corresponding to the first 4 Ts of the termination signal were considered as part of the transcripts; for non-canonical terminators, all the nucleotides of the terminator were considered as incorporated into the RNA. The underlined values are those for which a closely corresponding transcript was detected in transcription gels.

Table 2. MIRs subjected to in vitro transcription analysis

MIR	Expression in cell lines[1]	Predicted length of primary transcript(s)[2]
MIR_dup2285 (chr16:22309780-22309939	GM12878, H1-hESC, HeLa-S3, HepG2, HUVEC, K562, NHEK	**124/131** (T_4), **223/230** (T_3GT_2), **243/250** (T_4), **354/361** (T_3CT, downstream of annotated MIR but within the cloned sequence).
MIR_dup3493 (chr1:34943459-34943727)	GM12878, H1-hESC, K562, NHEK	**177** (TAT_3), **213** (TAT_3), **277** (T_2AT_2). Expected transcripts originating from terminators more downstream of the annotated MIR but within the cloned sequence: **305** (T_2AT_2), **365** (T_2AT_3), **393** (T_4).
MIRB_dup5848 (chr:71762977-71763215)	H1-hESC, HepG2, NHEK	**119** (TCT_3), **256** (T_3CT), **358** (T_5)
MIRC_dup2189 (chr14:89445565-89445634)	H1-hESC**,** K562, NHEK	**137** (T_3AT) and **140** (T_4), **209** (T_3GT) l, **250** (T_5)

[1]The column lists, for each MIR element, the cell lines in which it was found to be expressed by ENCODE RNA-seq data analysis.

[2]The reported transcript lengths were calculated by assuming as TSS the A or G residue closest to the position 12 bp upstream of the A box, by analogy with *Alus* (see Table 1). To estimate the 3' end of the transcript, both canonical (Tn with n≥4) and non-canonical T-rich Pol III terminators were considered both within and downstream of MIR body sequence (indicated in parentheses after the transcript length); for canonical terminators, the 4 Us corresponding to the first 4 Ts of the termination signal were considered as part of the transcripts; for non-canonical terminators, all the nucleotides of the terminator were considered as incorporated into the RNA. The underlined values are those for which one or more closely corresponding transcripts were detected in transcription gels. In the case of MIR_dup2285, for which two possible A boxes could drive transcription, the expected lengths of both putative alternative transcripts are indicated.

4. Results

4.1. ENCODE: *Alus*

4.1.1. A bioinformatic pipeline for the identification of transcriptionally active *Alu* loci from RNA-Seq datasets

The availability of RNA-Seq datasets for several human cell lines and tissues offers an unprecedented opportunity to identify individual, transcriptionally active *Alu* loci from the analysis of raw sequence reads. To this end, it is important to take into account the computational challenges posed by transcripts arising from repetitive elements, in particular the possible occurrence of multireads (i.e. reads aligning to multiple positions on the reference genome) (81). The RNA-Seq datasets we selected for our search are part of those established for the most recent ENCODE project attempt to define the landscape of transcription in human cells, and are all comprised of 76 nt-long paired end RNA-seq reads (91). In particular, we analyzed whole cell long RNA-seq data (polyA+ and polyA-) from ENCODE/Cold Spring Harbor Lab (http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeCshlLongRnaSeq) for the following cell lines: GM12878 (lymphoblastoid cells), H1-hESC (human embryonic stem cells), K562 (chronic myelogenous leukemia cells), HeLa-S3 (cervical carcinoma), HepG2 (hepatocellular carcinoma), HUVEC (umbilical vein endothelial cells), NHEK (epidermal keratinocytes). We considered in our analysis *Alu* elements differing in their location and possible mode of expression, in particular: (i) intergenic/antisense *Alus*, comprising both *Alus* that are not hosted in any annotated protein-coding or lincRNA gene and *Alus* that map to introns or exons of annotated genes, but do so in an antisense orientation (intergenic and antisense *Alus* were grouped together as they are both expected to be transcribed by Pol III as independent transcription units); (ii) *Alus* fully contained within introns of protein-coding or lincRNA genes in a sense orientation; (iii) *Alus* located within 5'UTR or 3'UTR of annotated protein-coding genes in a sense orientation; iv) all other cases, including *Alu* RNAs fully or partially mapping to exons in a sense

orientation. For groups (ii)-(iv), *Alu* RNA synthesis should in principle occur mostly as part of Pol II-dependent transcription of the host transcription unit, producing primary or mature mRNA/lincRNA transcripts carrying embedded *Alu* RNA. These different possibilities for *Alu* location with respect to other transcription units are illustrated in Figure12.

Figure 10 provides a schematic representation of the pipeline we devised to map ENCODE sequence reads to human *Alu* collections. Our search strategy displays two main features introduced to ensure as much as possible the identification of genuine *Alu* transcripts (i.e. transcripts whose start, end and sequence closely match those expected from Pol III-dependent transcription of a particular annotated *Alu* element). The first such feature, aimed at avoiding ambiguous mapping due to *Alu* repetitive nature, is the reconstruction of base-resolution expression profiles of individual *Alus* based exclusively on sequence reads that do not map to any other genomic location. This task was accomplished through the TopHat aligner, and was facilitated by the paired-end nature of ENCODE RNA-seq data, allowing unique mapping not simply on the basis of the sequence of individual reads, but also of the combination of sequence and colocalization of the two 76-nt mates in the same read pair (see Materials and Methods and Supplementary Materials for details). Such a 'unique' alignment strategy (81) might lead to underestimate the number of expressed *Alus* (as well as general *Alu* expression levels); in particular, expressed *Alus* present in multiple identical copies would be overlooked. To take this possible limitation into account, a parallel and more permissive analysis was also conducted in which each individual read mapping to more than one site was not discarded, but randomly attributed to one of the matching genomic sites ['best match' alignment strategy in ref. (81)]. Most of the data presented in this study were based on unique alignment, as the unambiguous identification of expression-positive *Alu* loci was our main task. The less stringent, 'best match' alignment was only employed for some analyses, which would have been compromised by the exclusion of reads that were not uniquely mappable (see below).

The second key feature of the search pipeline is a filter step which, by imposing a requirement for significantly lower read densities to the flanking regions immediately upstream and downstream of each *Alu* element, systematically excludes *Alu* RNA sequences that are part of longer, Pol II-synthesized transcripts. This filter is also aimed at excluding *Alu* RNAs that are part of Pol II transcript trailers

extending downstream of annotated 3'UTRs. A shortcut for the elimination of embedded *Alu* RNA could have been the *a priori* exclusion, from the reference *Alu* dataset, of any *Alu* mapping in a RefSeq gene in a sense orientation. In this case, however, a number of potentially interesting cases might have been overlooked. Indeed, the Pol III machinery might in principle also act on *Alus* embedded in Pol II gene introns or UTRs, to produce free (not embedded) *Alu* RNAs. As a further possibility, intron-located *Alu* RNAs might be released from the host intron RNA through intron processing, as it occurs for intron-derived microRNAs or snoRNAs (92, 93). Both nested Pol III transcription units and *Alu* RNA maturation from introns would generate *Alu* RNAs passing the final filter step for independently expressed *Alu* transcripts in our search pipeline. For each dataset our search thus considered, as potential transcript sources, all *Alus* (either complete or incomplete), while maintaining a distinction among: (i) intergenic/antisense *Alus*; (ii) intronic sense-oriented *Alus*; (iii) 5'/3' UTR-embedded, sense-oriented *Alus*; (iv) partially or fully exonic sense-oriented *Alus*. We chose to consider as expressed those *Alus* with 10 or more uniquely mapped read pairs (see the Materials and methods section for the rationale for this choice).

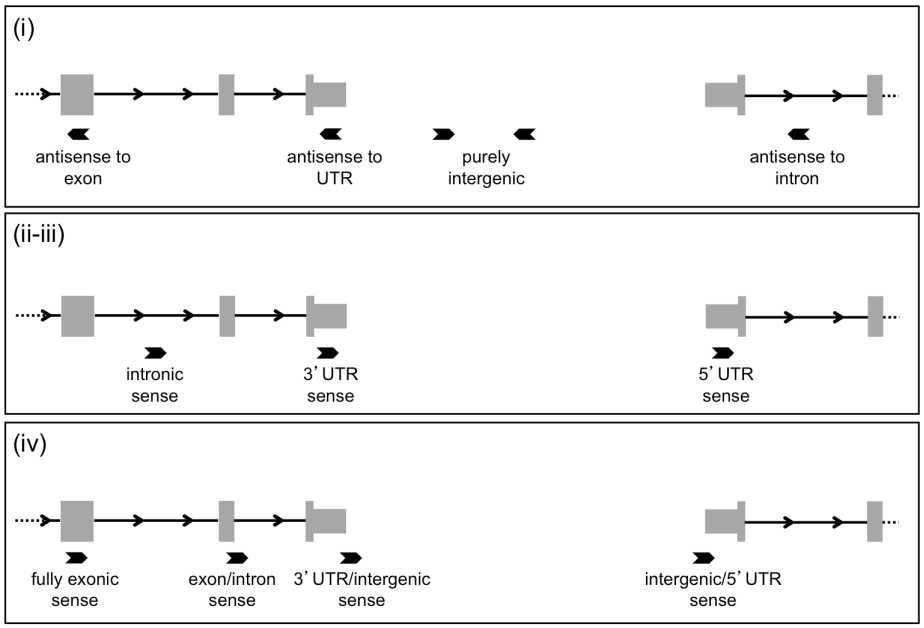


Figure 12. Possible localizations of *Alu* elements with respect to other transcription units: i) intergenic/antisense, comprising purely intergenic *Alus* as well as *Alus* which are not included in longer transcription units on the same strand, but overlap in antisense orientation to transcription units located on the opposite strand; ii-iii) gene-hosted, comprising *Alus* fully contained within introns or UTRs of protein-coding or lincRNA genes in a sense orientation; (iv) all other cases, including *Alu* RNAs fully or partially mapping to exons, or partially mapping to UTRs, in a sense orientation.

4.1.2. General features of *Alu* transcriptomes emerging from ENCODE RNA-seq data analysis

The full list of *Alus* identified as expressed by our search algorithm is provided in Supplementary Table S1. We observed that more *Alu* RNAs are recovered in the polyA- than in the polyA+ fraction of cellular RNA. In detail, 394 and 968 individual *Alu* RNAs were collectively identified in polyA+ and polyA- fractions, respectively; among these, *Alu* RNAs originating from 67 *Alu* loci were found in both polyA+ and polyA- fractions. Therefore, even though *Alu* RNAs contain an intermediate A-rich spacer and a 3'-terminal poly(A) or A-rich tract which might

facilitate their inclusion into poly(A)-containing cellular RNA fractions (94), the A tracts of the majority of them are not sufficiently long for inclusion in polyA+ RNA
A preliminary survey of base-resolution expression profiles of individual *Alus* characterized by different locations suggested that the search pipeline was very effective in identifying intergenic *Alus* autonomously expressed from their Pol III promoter. Several cases of gene-hosted, sense-oriented *Alus*, whose transcripts appear to accumulate independently from host gene expression, could also be identified. Among *Alu* RNAs mapping to gene-hosted elements, however, the final filter step did not appear to be completely effective in excluding spurious *Alu* RNA sequences that are probably part of longer intronic or messenger RNA molecules. Such ambiguous signals were frequently observed in correspondence of incomplete *Alu* elements, whose base-resolution profiles, allowing them to pass the filter test, could be explained by the presence of transcribed non-*Alu* sequences flanking the incomplete *Alu* upstream and/or downstream, but likely deriving from Pol II transcription of the host gene. For these reasons, we decided to mainly focus on the expression of intergenic/antisense *Alus*, while a few examples of gene-hosted *Alus* will be addressed later in the Results section.
As summarized in Table 3, each of the cell lines expressed a limited number of *Alu* elements (ranging from 149 in the case of Gm12878 cells to 425 in the case of HepG2 cells). Of the whole set of 1295 expression-positive *Alu* loci, about 30% displayed an intergenic/antisense location (including both purely intergenic *Alus* and *Alus* overlapping with annotated genes in an antisense orientation). Among expression-positive intergenic/antisense *Alus*, a significant percentage (~22%) actually mapped in antisense orientation to introns of annotated, Pol II-transcribed genes (including 10 lincRNA genes). A consistent fraction (20%) of the expression-positive intergenic/antisense *Alus* were found to be expressed in more than one cell line. On average, ~40% of intergenic/antisense *Alus* expressed in a cell line were also expressed in at least one other cell line. Given the extremely high number of genomic *Alus* which could in principle be expressed, on the order of hundreds of thousands, such a marked sharing of actually expressed *Alus* among different cell lines, each of which expresses no more than 0.1% of total *Alus*, points to the existence of a tiny subset of "transcription-prone" *Alu* elements, within which cell type-specific differences in *Alu* expression profiles can be established. *Alus* have been classified into three main subfamilies, called *Alu*J, *Alus* and *Alu*Y, and it has been

proposed that *Alu*Y elements, being the youngest evolutionarily, and thus the less degenerated in sequence, might represent the most transcriptionally active subfamily, in agreement with the observation that the only known *Alu* elements currently active in retrotransposition in the human genome belong to the *Alu*Y subfamily (95). We thus asked whether a higher tendency to be expressed could be put in light for *Alu*Y with respect to *Alu*J and *Alus* subfamilies. As summarized in Table 4, no significant over-representation of any particular *Alu* subfamily within the set of expressed *Alus* was observed when intergenic *Alus* only were considered, while *Alu*Y, somehow unexpectedly, appeared to be slightly under-represented when the full set of expression-positive *Alus* was considered. Since the above data are based on uniquely mapped reads, the younger AluY subfamily, whose individual members tend to be more homogeneous in sequence, could be under-represented among expression-positive *Alus* simply because of a wider exclusion of the corresponding reads as non-uniquely mapped. To avoid such a possible bias, we interrogated for *Alu* subfamily representation an *Alu* expression dataset generated through a variant of our search pipeline in which the TopHat aligner, through the "-g1" setting, distributes multireads randomly across equally good loci (see Supplementary Materials). We reasoned that, in this way, most *Alu*Y multireads, discarded in the unique alignment, would be attributed to members of the same subfamily. As shown in Table 4, the *Alu*Y under-representation was less marked in this case, thus leading to conclude that no differential expression of specific *Alu* subfamilies is put in evidence by our analysis.

Table 3. Statistics of expression-positive *Alu* elements in selected cell lines

Cell line	Total *Alus*[1]	Intergenic/antisense	Intergenic/antisense shared[2]	Antisense to introns[3]
GM12878	149	48	21	13
H1-hESC	257	92	28	12
HeLa S3	276	44	32	7
HepG2	425	88	31	19
HUVEC	326	36	18	4
K562	154	71	33	20
NHEK	231	130	38	34
ALL[4]	1295	386	78	87

[1] For each cell line, the column reports the number of *Alus* considered as autonomously expressed in both ENCODE RNA-seq replicates.
[2] For each cell line, the column reports the number of intergenic *Alus* that are also expressed in one or more different cell lines.
[3] Reported in this column are the numbers of intergenic *Alus* mapping with an antisense orientation to introns of both protein-coding and lncRNA genes.
[4] The numbers in this raw refer to individual *Alus* expressed in one or more cell lines.

Table 4. Subfamily distribution of expression-positive *Alus*

Alu subfamily	Total genomic[1]	Expressed genomic[1]	Total intergenic/ antisense[1]	Expressed intergenic/ antisense[1] (copy number)	Expressed intergenic/ antisense[2] (read count)
S	675428 (60%)	735 (57%)	513048 (60%)	219 (57%)	62%
J	307612 (27%)	479 (37%)	225907 (27%)	112 (29%)	27%
Y	140707 (13%)	81 (6%)	107922 (13%)	55 (14%)	11%
TOTAL	1123747	1295	846877	386	100%

[1] Reported are the absolute copy numbers and (in parentheses) the percentages of *Alus* of each sub- family considered relative to (from left to right): the total set of genomic *Alus* ("Total genomic"); the set of *Alus* found to be expression-positive in one or more cell line ("Expressed genomic"); the total set of intergenic/antisense *Alus*; the set of intergenic/antisense *Alus* found to be expression-positive in one or more cell lines;
[2] This column refers to the dataset of expressed intergenic/antisense *Alus* generated through a variant of the search pipeline in which the TopHat aligner, through the "-g1" setting, distributes multireads randomly across equally good loci.

Of the 1295 unique putative *Alu* transcripts discovered by our bioinformatic pipeline, approximately 9% are expressed in at least 3 cell lines, while ~75% of them turned out to be cell type-specific (see Supplementary Table S1). Thus, despite the tiny fraction (0,01%) of *Alu* loci found to be expressed among the ~1.1 million *Alus* in the human genome, these results suggest that Pol III-transcribed *Alu* RNAs derive from a small subset (~100) of ubiquitously expressed *Alu* elements, and from a larger subset (~1000) that tends to vary by cell type, state, growth conditions. This observation is in agreement with the results of recent human transcriptome analyses, showing a marked cell line specificity as the main feature of RNAs transcribed from repeated regions, including LINEs and SINEs [30]. A heatmap visualization of cell lineage-specific *Alu* expression is reported in Figure 13.

Figure 13. Cell lineage-specificity of *Alu* expression. Shown is the heatmap of expression-positive *Alus* from the indicated cell lines, sorted on the basis of cell-line specific expression, displaying ubiquitously expressed/non-specific *Alus* (left), and tissue-specific *Alus* (right).

4.1.3. Survey of expressed intergenic *Alus* according to location and base-resolution expression profiles

The inspection of individual expression profiles reconstructed through our analysis for intergenic *Alus* revealed different types of profile that deserve circumstantial examination. In particular, profiles were observed which can be roughly summarized as: whole *Alu*, left-monomer, right-monomer coverage.

Figure 14 shows examples of the occurrence of these expression profiles for both purely intergenic and intron-antisense *Alus*. For the *Alu*Sg reported in Figure 14A, found to be expressed in three different cell lines (H1hESC, K562 and NHEK), a complete and precise coverage of the *Alu* by uniquely mapping sequence reads was observed. An inspection of its sequence revealed that this *Alu* possesses canonical A- and B-boxes, as well as a Pol III termination signal located ~20 bp downstream of

the 3' poly(dA) tail, thus suggesting that Pol III transcription of this intergenic transcription unit generates a specific, ~300 nt-long *Alu* RNA. Figures 14B and 14C show typical examples of expression profiles corresponding to truncated *Alu* transcripts. Frequently, sequence reads tended to cover either the left or the right monomer of a complete *Alu* element. For the *Alu*Y of Figure 14B, sequence reads precisely covered the left monomer sequence, up to the short A-rich region (A_5TACA_6) separating the two *Alu* monomers. Given the absence of Pol III termination signals within the body of this *Alu*, the short transcript likely belongs to the previously reported family of small cytoplasmic (sc) *Alu* RNAs (96), being generated by processing of a full-length primary *Alu* transcript (97). Truncated *Alu* transcripts like the one reported for the *Alu*Sc in Figure 14C are more difficult to interpret based on our current understanding. In this case the transcript appears to start just downstream of the internal A-rich region, suggesting that right monomer *Alu* RNA fragments might also be generated through processing of full-length precursors. Incomplete coverage of some *Alus* might in principle be due to the fact that these *Alus* possess sequence tracts (corresponding to the uncovered regions) that are identically repeated at other genomic locations, such that mapping reads would be non-unique and thus discarded. To explore this possibility, we looked at the coverage profiles obtained for some of these *Alus* (those in Figures 14B and 14C) using the TopHat bam file generated with default settings, and thus reporting up to 20 alignments for multi-mapped reads. We still observed the same incomplete coverage for all of these *Alus* (Supplementary Figure S1). Incomplete expression profiles are thus unlikely to be due to multimapping issues, as they are not appreciably changed by multiread-permissive alignment. Furthermore, the fact that the same partial coverage profiles were also observed with STAR alignment (also shown in Supplementary Figure S1) argues against partial coverage being an aligner artefact. In Figure 14D, the whole *Alu*Y element (mapping with antisense orientation to the second intron of the COL4A1 gene) is covered by sequence reads all along its extension, but with a double-humped profile in which two peaks are approximately centred on the left and right monomer of the *Alu* element. This type of profile was much more frequently observed in our analyses than the more continuous type of profile such as the one shown in Figure 14A. As a tentative explanation, we reasoned that, since a full-length *Alu* transcript is on average 300-nt long, the post-fragmentation selection of 200 bp fragments during RNA-seq library

preparation (91) should produce a relative enrichment in fragments containing either the 5' or the 3' end of the *Alu* cDNA. Sequencing of the 3' and 5' ends of such cDNA fragments would lead to an under-representation of the central part of the transcript, and thus to the generation of two-humped base-resolution profiles.

We also identified cases of incomplete *Alu* elements (*Alu* monomers) whose corresponding RNAs extend upstream or downstream of the annotated *Alu* monomer. Figure 14E shows the case of a 150-bp long, right *Alu*Sx monomer (mapping with an antisense orientation to the first intron of HRH1, just upstream of the next-to-last exon), whose sequence read coverage extend ~60 bp upstream, delineating a transcription unit starting upstream of the *Alu* monomer, within an *Alu*-unrelated region, and including the *Alu* right monomer as the downstream moiety of the transcript. A complementary example of an *Alu* left monomer being part of a longer transcription unit extending downstream is reported in Figure 14F, showing the expression profile of a purely intergenic *Alu*Sg7. Here a ~120-bp left monomer containing A- and B-boxes appears to direct the synthesis of a transcript ending approximately 180 bp downstream, at a position which is only ~400 bp upstream of the TSS of the SEC61G gene. Through parallel analysis of ENCODE ChIP-seq data of Pol III components, we noted the existence of Pol III and TFIIIC association peaks precisely mapping to this *Alu*, an observation supporting the conclusion that it constitutes a *bona fide* Pol III transcription unit. (A more exhaustive account of parallel analysis of ENCODE ChIP-seq data will be provided below). Through the "-g1" variant of our search algorithm, attributing multireads randomly to one of the hits, we observed another interesting case of an *Alu* left monomer directing the transcription of a longer transcription unit (see below "Identification of a novel *Alu*Ya5-derived Pol III transcript"). Expression of *Alu* monomers is thus likely to be more frequent than commonly thought, in agreement with the observation of recent *Alu* monomer insertions, some of which generated through retroposition (98). Interestingly we observed, as a general trend for expression-positive *Alu* monomers, that transcripts mapping to left and right *Alu* monomers extend downstream and upstream of the monomer, respectively, in agreement with the fact that *Alu* left monomers generally contain a functional Pol III promoter, able to direct transcription of the monomer itself followed by downstream sequences until a Pol III terminator is encountered, while *Alu* right

monomers do not contain a Pol III promoter and thus their expression requires incorporation into an upstream initiated transcript.

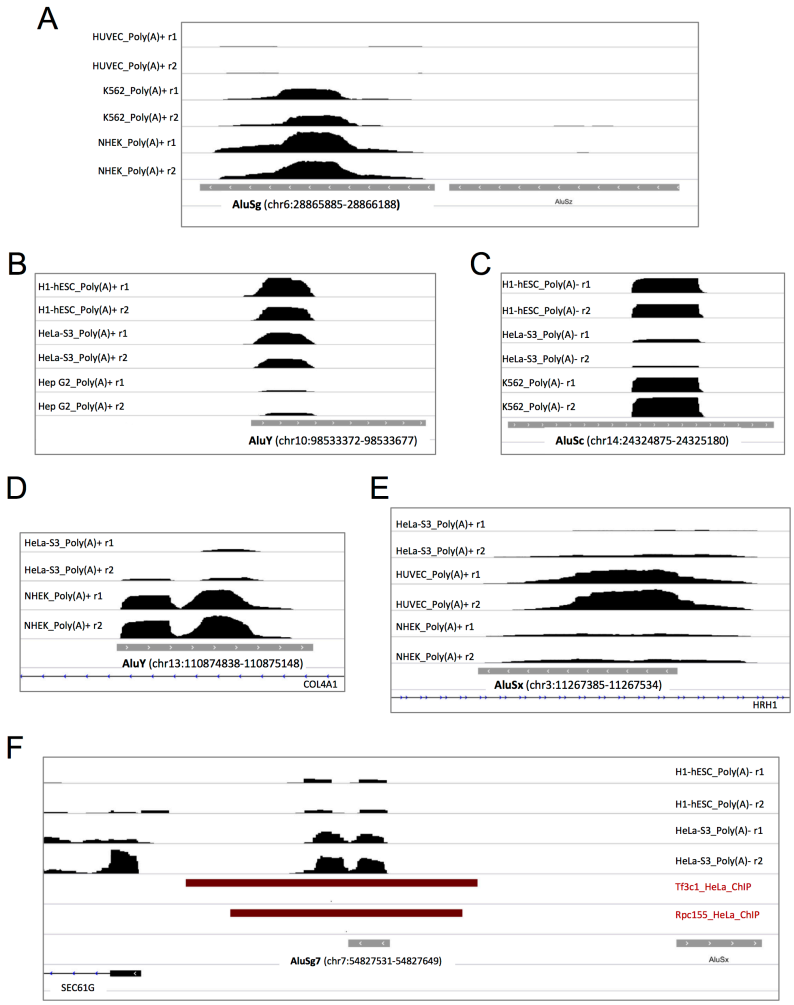


Figure 14. Base-resolution expression profiles for six representative *Alus* of the intergenic/antisense type. Panels *A*-*C* and *F* refer to purely intergenic *Alus*, panels *D* and *E* to two antisense *Alus*. Shown are the Integrative Genomics Viewer (IGV; http://www.broadinstitute.org/igv/home) visualizations of RNA-seq stranded expression profiles (in bigwig format) around *Alu* loci in the cell lines indicated either on the left (*A*-*E*) or on the right (*F*) of each panel. r1 and r2 indicate the two independent replicates found in ENCODE data. The orientation and chromosomal coordinates of each *Alu*, as well as the overlapping (antisense) or nearby RefSeq genes, are indicated in each panel. The dark red bars in panel *F* indicate regions associated to either TFIIIC (Tf3c1 track) or Pol III (Rpc155 track) in HeLa cells as derived from ENCODE ChIP-seq data.

4.1.4. Evidence for independent expression of gene-hosted, sense-oriented *Alus*

Even though *Alus* located within intron or exons (including UTRs) of Pol II-transcribed genes are expected to be mostly transcribed as part of longer Pol II transcripts, we addressed the possibility that a few of them might be transcribed as autonomous Pol III transcription units or, more generally, that the corresponding *Alu* RNAs might accumulate to a detectable extent independently from host gene expression. The final filter step of our search algorithm is devised to produce an enrichment of such *Alu* RNA species, as it imposes a strong reduction in the number of sequence reads mapping to regions flanking the gene-hosted *Alus*, thus favouring isolated expression signals centred on *Alu* elements. By inspecting the profiles of many gene-hosted (especially intron-hosted) *Alus* that had been identified as expression-positive in our search, we confirmed the presence of *Alu*-centred expression signals as expected on the basis of our filter step; the *Alu* peaks, however, were frequently preceded and/or followed by expression peaks mapping to *Alu*-less surrounding regions, thus suggesting the possibility that *Alu* signals, as well as the surrounding signals, might represent fragments of longer intron RNAs. In a limited number of cases, however, *Alu* expression profiles were suggestive of the presence of autonomous *Alu* transcription units. One such case is illustrated in Figure 15A, showing the base-resolution expression profile of an *Alu*Sx1 located, in a sense orientation, within the first intron of SRGAP2, a gene involved in human brain development and evolution (99). The *Alu*Sx1 is followed immediately downstream by an *Alu*Sp with the same orientation, to which a few sequence reads also map. The left monomer of the *Alu*Sx1 has canonical A- and B-boxes, but the first potential Pol III terminator is located downstream of the *Alu*Sp, thus suggesting that these two *Alus* might be transcribed into a dimeric *Alu* primary transcript. A similar situation is illustrated by the example in Figure 15B, reporting the profile of an intronic sense-oriented *Alu*Y located between exons 9 and 10 of ZC3H3 gene. This *Alu* Y is endowed with A- and B-boxes, and even if there is no recognizable Pol III termination signal separating the *Alu*Y from the *Alu*Sq located immediately downstream with the same orientation, transcription appears to terminate just downstream of the first *Alu*, given the absence of sequence read coverage of the second *Alu*. However, through parallel analysis of ENCODE ChIP-

seq data of Pol III factors, we noted that Pol III (and TFIIIC) appear to be associated with a region encompassing both *Alu*Y and the downstream *Alu*Sq, as if both were part of the same transcription unit. An intriguing example of independent accumulation of intronic *Alu* RNA is provided by the *Alu*Jb located within the intron separating exons 35 and 36 of USP34 (Figure 15C). The expression levels of this left *Alu* monomer (whose transcripts extend downstream by ~70 bases) appear to be inversely correlated with the levels of exon 37 expression in the different cell lines, suggesting mutual expression interference. A few cases of independently expressed *Alus* located within lincRNA gene introns were also observed. One of them is illustrated in Figure 15D, showing the base-resolution profiles of an *Alu*Y hosted in a sense orientation between exons 4 and 5 of lincRNA gene TCONS_I2_00015350 on chromosome 2. ChIP-detected association of Pol III and TFIIIC with this *Alu* locus further argues that it is a genuine Pol III transcription unit. Finally, as exemplified in Figure 15E, 3' UTRs can also host sense-oriented *Alus* whose transcripts accumulate independently from the corresponding mRNA.

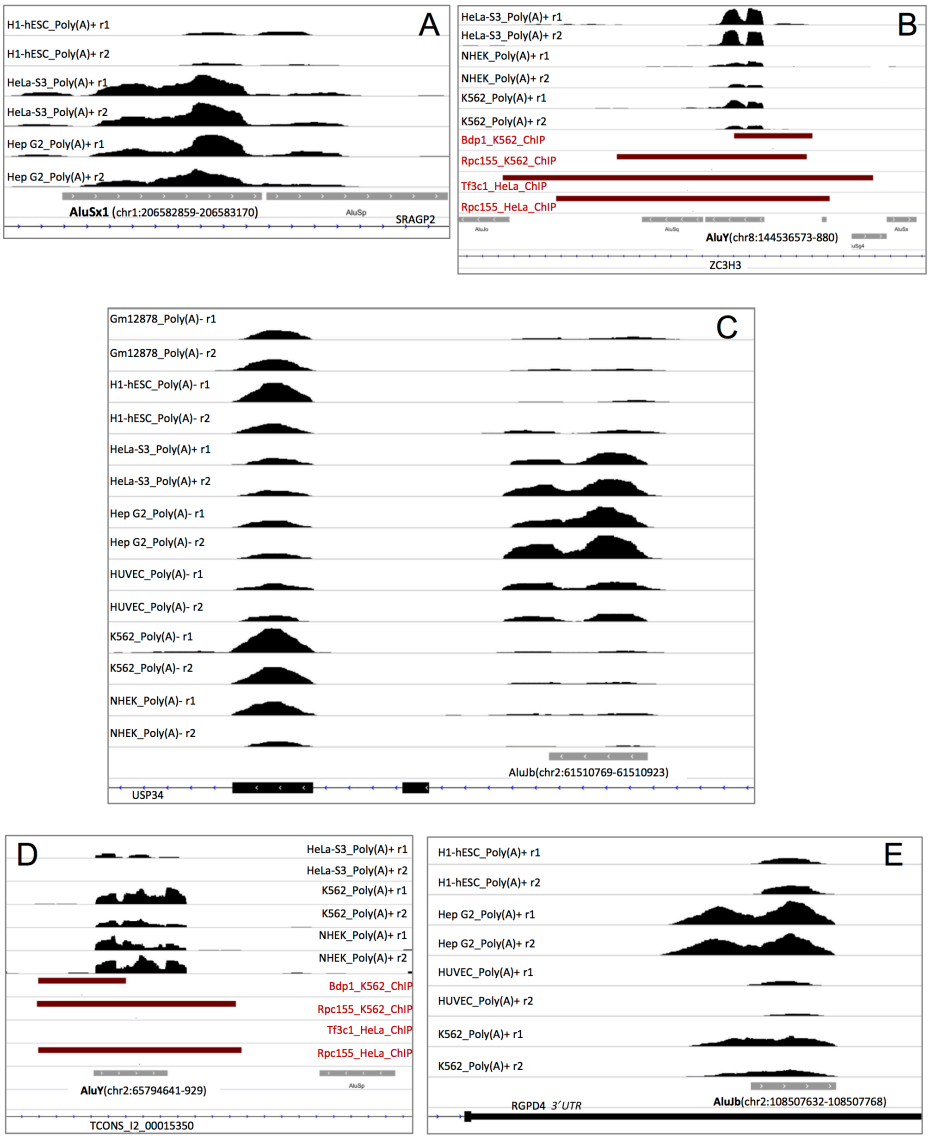


Figure 15. Base-resolution expression profiles for five representative gene-hosted, sense-oriented *Alus*. Panels *A*-*C* refer to *Alus* hosted within introns of RefSeq genes, panel *D* to a 3'UTR-hosted *Alu*, panel *E* to an *Alu* hosted within a a lincRNA gene intron. Shown are the IGV visualizations of RNA-seq stranded expression profiles (in bigwig format) around *Alu* loci in the cell lines indicated either on the left (*A*-*D*) or on the right (*E*) of each panel. r1 and r2 tracks refer to the two independent replicates found in ENCODE data. The orientation and chromosomal coordinates of each *Alu*, as well as the host RefSeq or lincRNA genes, are indicated in each panel. The dark red bars in panels *B* and *F* identify regions associated to the indicated Pol III transcription component (Bdp1, Tf3c1 or Rpc155) in either K562 or HeLa cells as derived from ENCODE ChIP-seq data.

4.1.5. Association of the Pol III machinery to expression-positive *Alus*

Several genome-wide association studies based on ChIP-seq approaches have been conducted in the last few years with the aim of producing complete inventories of Pol III-transcribed genes (reviewed in (24)). Each of these studies identified a variable (generally small) number of *Alus* associated to the Pol III machinery. In a recent study, an integrated, comparative evaluation of Pol III-associated *Alus* was carried out through a synopsis of several ChIP-seq studies (70). We asked whether there is any significant overlap between the set of *Alus* identified as expressed in our analysis and the Pol III-associated *Alus* in ChIP-seq studies. To address this point we took advantage of the availability, within the ENCODE data, of ChIP-seq datasets, relative to both K562 and HeLa-S3 cell lines, for key components of the Pol III transcription machinery: Bdp1 and Brf1 (components of TFIIIB), Rpc155 and TFIIIC110 (subunits of RNA polymerase III and TFIIIC, respectively). Supplementary Table S2 lists the expression-positive *Alus* that are also associated to Pol III components in HeLa and K562 cells. In HeLa cells, 15 out of 276 expression-positive *Alus* (~6%) were found among those associated to one or more components of the Pol III machinery in the ENCODE datasets. When the comparison was restricted to the 44 intergenic *Alus* detected as expressed in HeLa cells, a much higher fraction of them (29%, 13 *Alus*) were also associated with the Pol III machinery, with 11 *Alus* being associated with at least two transcription components and 8 with three components representing the whole machinery (TFIIIB, TFIIIC, Pol III). P-values for association of Bdp1, TFIIIC110 and Rpc155 with intergenic expressed *Alus* (*vs.* the whole set of intergenic *Alus*) were all $< 10^{-14}$. Similarly, when K562 cells were considered, a significant percentage of expression-positive *Alus* was Pol III-associated and most strikingly, of the 71 intergenic expression-positive *Alus* in these cells, 31 (corresponding to 44%) were found associated with at least one component of the Pol III machinery. P-values for association of Bdp1, TFIIIC110 and Rpc155 with intergenic expressed *Alus* (*vs.* the whole set of intergenic *Alus*) in K562 cells were $< 10^{-15}$ (Supplementary Table S2). Specifically, of the intergenic *Alus* whose expression profiles were shown in Figure 14, three were found to be associated to either Pol III (chr13:110874838-110875148, panel D) or Pol III and TFIIIC (chr6:28865885-28866188, panel A; chr7:54827531-54827649, panel F). Interestingly, a few intronic sense-oriented *Alus* identified as

autonomously expressed were also found to be associated with components of the Pol III machinery; among them were those whose profiles are shown in Figure 15B (chr8:144536573-880) and 15D (chr2:65794641-929). Altogether these findings confirm the effectiveness of our *Alu* RNA detection procedure, especially in the case of intergenic *Alus* but also for intron-hosted elements, and suggest the existence, in each cell type, of a very small and specific subset of individually trackable, transcription-prone *Alus*. We noted that only four *Alu* elements were found to be expressed and Pol III-associated in both K562 and HeLa cell lines (chr1:61523296-61523586; chr10:5895538-5895651; chr1:28672563-28672802; chr8:144536572-144536880), suggesting a high plasticity of the *Alu* transcriptome.

4.1.6. Identification of a novel *Alu*Ya5-derived Pol III transcript

In parallel with a stringent search procedure based on a 'unique alignment' strategy, we also applied to ENCODE RNA-seq datasets a 'best match' alignment strategy (81), in which multireads are attributed randomly to one of the hits, with the aim of detecting expressed *Alus* whose presence in multiple identical copies in the genome would prevent their identification as expression-positive in the unique alignment strategy. In this case, the analysis was restricted to intergenic/antisense *Alus*. As expected, a significantly higher number of intergenic/antisense *Alu* elements were identified with respect to 'unique alignment' search (705 versus 386). Through systematic inspection of *Alus* found as expression-positive in at least three cell lines, we discovered multiple (~20) almost identical copies of an *Alu*Ya5 left monomer, encompassed within the recently described snaR A/C and snaR A/B/D clusters on the q-arm of chromosome 19 (100). Base-resolution expression profiles of these *Alu*Ya5 elements suggest that transcription initiates at the *Alu* monomer and continues downstream of it, in a 3'-flanking region whose sequence is *Alu*-unrelated. In this respect these transcription units, hereafter referred to as Ya5-lm (for left monomer of *Alu* Ya5), resemble the BC200 RNA gene, which can also be described as a transcriptionally active *Alu* element consisting of an upstream monomeric *Alu* repeat followed by a non-repetitive domain (101). The base-resolution expression profile of three clustered Ya5-lm elements is shown in Figure 16A, in which their expression can be directly compared with the Pol III-dependent expression of

interposed SNAR-A3 elements. Reported in Figure 16B are the sequence and the general organization of Ya5-lm. The upstream *Alu*Ya5 monomer contains typical *Alu* A- and B-boxes (with the A-box differing from canonical tRNA A-box for a C instead of G at the last position; (27)), and ends with an A-rich motif. Downstream of this motif the sequence of the Ya5-lm transcription unit diverges from consensus *Alu* sequence (Figure 16C). The first potential Pol III termination signal (TTTT) starts at position +260, almost exactly corresponding to the end of sequence read coverage. The snaR genes on chromosome 19 are arrayed in two large inverted regions of tandem repeats, with the two clusters (A/C and A/B/D) separated by a 2-Mb region (100, 102). We found that these two clusters contain 11 and 10 copies of Ya5-lm, respectively. In both clusters, all Ya5-lm, separated from each other by ~5300 bp, have the same orientation as snaR genes, and each of them is separated by ~1800 bp and ~3300 bp from the upstream and downstream snaR gene, respectively. As the Ya5-lm copies on chromosome 19 are almost identical, it is difficult to specifically attribute to one or more of them the mapping sequence reads. Nevertheless, since the non-repetitive sequence domain downstream of *Alu*Ya5 monomer is not found at any other locus in the genome, there is no doubt that one or more of these genes are transcribed to produce a novel type of *Alu*-derived Pol III transcript. In support to this conclusion are ChIP-seq data from Pol III genome-wide location studies available at ENCODE. In one of them, Pol III-associated loci in K562 cells were identified through ChIP-seq using an antiserum against the Pol III largest subunit Rpc155 (103). Analysis of Rpc155 ChIP signals revealed a peak precisely overlapping with the *Alu*Ya5 identified by the coordinates chr19:50640453-50640584, and by transcript coverage, in both HeLa and K562 cells (Figure 16D). It is thus likely that only one (or a small subset) of the Ya5-lm elements on chromosome 19 are transcriptionally active. The attribution of multireads randomly to one of the hits in the 'best match' alignment strategy explains why all Ya5-lm copies are covered by sequence reads (as exemplified in Figure 16A).

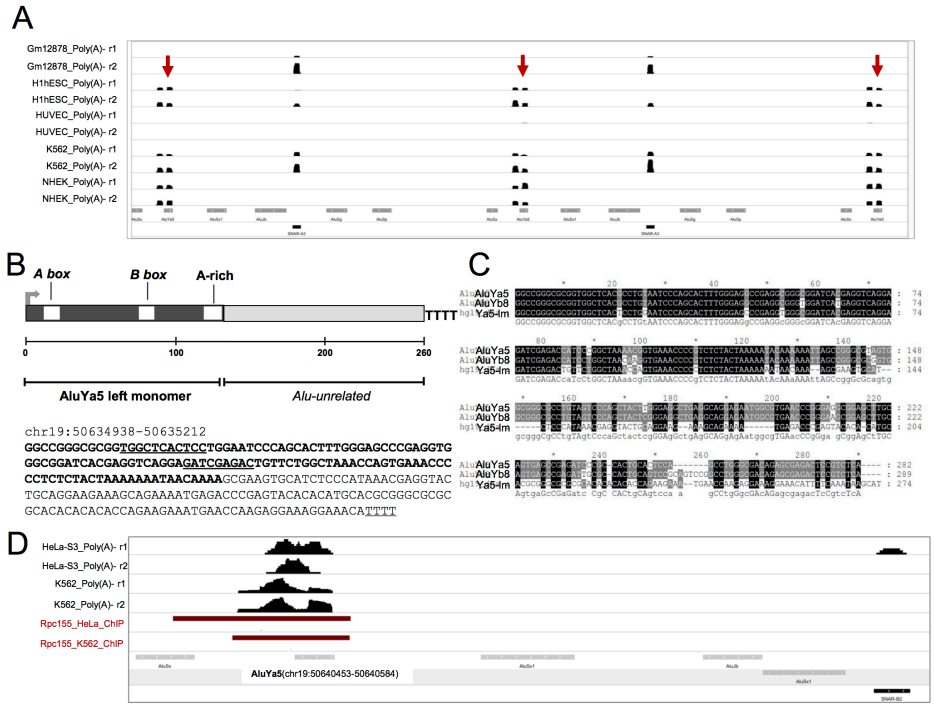


Figure 16. Novel *Alu*Ya5-derived transcription units associated to snaR clusters. (*A*) Genome browser visualization of RNA-seq stranded expression profiles of three *Alu*Ya5-derived transcription units (Ya5-lm, indicated by red arrows) within the snaR A/C/D cluster on chromosome 19 (100). (*B*) Transcription unit architecture and sequence of a Ya5-lm repeat (coordinates in parentheses). (*C*) Sequence alignment of Ya5-lm with Repbase reference sequences for *Alu*Ya5 and *Alu*Yb8. (*D*) Genome browser visualizations of RNA-seq stranded expression profiles around the Ya5-lm element represented in panel *B*, in the cell lines indicated on the left. The dark red bars identify regions associated to Pol III (Rpc155 subunit) in either K562 or HeLa cells as derived from ENCODE ChIP-seq data.

4.1.7. In vitro transcription analysis of expressed and silent *Alu* elements

The ability to detect *in vivo* expression of individual *Alu* elements prompted us to verify whether *Alus* with different expression levels can also be differentiated for their *in vitro* transcription behaviour. To this end, we focused on a small subset of *Alu* loci, representative of different types of expression profiles based on the analyzed RNA-seq datasets. These loci are listed in Table 1. One of them, *Alu*Sq2_chr1:61523296-61523586, appears to be expressed in five different cell lines (H1-hESC, HeLa-S3, Hep G2, K562, NHEK), and was also found associated to the

Pol III machinery in both HeLa and K562 cells (see Supplementary Table S2). This *Alu* was thus chosen as representative of ubiquitously expressed *Alus*. Moreover, this *Alu* is peculiar in sequence as it lacks the internal A-rich motif A_5TACA_6, which is replaced by A_3G. Two other loci, *Alu*Y_chr10:69524852-69525156 and *Alu*Sp_chr17:4295121-4295437, are expressed above our chosen threshold in only one out of seven cell lines [NHEK and K562 cells, respectively; but lower levels of expression were detectable in other cell types; interestingly, *Alu*Y_chr10 is among the few *Alus* identified as expressed in this study that were also among the candidate source *Alus* in a recent analysis (70)]. Three loci (*Alu*Sx1_chr3:139109300-139109588, *Alu* Y_chr7:73761603-73761897, *Alu*Y-_chr10:103929453-103929749) were found to be expressed in a somewhat sporadical manner (i.e. in no more than two cell lines and in one replicate only); however, based on ENCODE ChIP-seq data, each of them is associated to one or more components of the Pol III machinery. The remaining three loci (*Alu*Sx_chr1:235531222-235531520, *Alu*Sx_chr10:12236879-12237173, *Alu*Y-_chr22:41932115-41932411) were not found to be detectably expressed by our analysis, even though the *Alu*Sx on chromosome 10 was Pol III-associated based on ENCODE ChIP-seq data. Five of these *Alus* have a purely intergenic location (i.e. they do not overlap with any other transcription unit in either antisense or sense orientation), while four of them are antisense with respect to introns of protein-coding genes. Interestingly one of them, *Alu*Y_chr22:41932115-41932411, maps in antisense orientation to intron 2 of POLR3H, coding for the 22.9-KDa subunit of RNA polymerase III (RPC8/RPC22.9), thus suggesting a possible role of this element in POLR3H gene regulation, as already proposed for a MIR elements located on the minus strand within the first intron of both human and mouse genes coding for the RPC5 subunit of Pol III (42, 44). The other selected antisense *Alus* map to: the first intron of TBCE (Tubulin Folding Cofactor E) gene (*Alu*Sx_chr1:235531222-235531520); the third intron of CLIP2 gene (*Alu*Y_chr7:73761603-73761897); the first intron of NUDT5 gene (*Alu*Sx_chr10:12236879-12237173). The 9 selected *Alu* elements were PCR-amplified from human genomic DNA, cloned into pGEM-T-easy vector, and tested for their ability to support efficient *in vitro* transcription using a HeLa cell nuclear extract. To verify that the observed transcripts were produced by the Pol III machinery, reactions were conducted in the presence of α-amanitin at a concentration (2 μg/ml)

known to completely inhibit RNA polymerase II activity, and transcription reactions were also programmed in parallel with a mutant version of each *Alu* element, in which the B box internal promoter element was mutationally inactivated. The results of *in vitro* transcription analysis are shown in Figure 17. Control transcription reactions were programmed with empty pGEM-T-easy plasmid (lanes 1, 11, 21) and the same vector carrying either (lane 2, 12, 22) a previously characterized, transcriptionally active *Alu* (*Alu*Sx1_chrX:24096144-24096441, producing a 372-nt transcript; Orioli, A. and Dieci, G., unpublished data) or (lane 3, 13, 23) a $tRNA^{Val}$(AAC) gene (TRNAV18, chr6) whose transcription produces three different primary transcripts (of 87, 112 and 142 nt) because of heterogeneous termination at one of three consecutive termination signals (31). Each of the tested *Alu* elements produced a well-defined pattern of transcription, in which the sizes of the longest and most abundant transcripts matched those predicted on the basis of sequence inspection of the Pol III termination signals, either canonical (a run of at least four Ts) or non-canonical (31), in the 3'-flanking region. The observed transcription efficiencies of all *Alus* were comparable (with the exception of *Alu*Sx_chr1 (lane 7) producing low levels of transcription products heterogeneous in size), indicating that their different tendency to be transcribed in cultured cells is not due to differences in *cis*-acting elements recognized by the basal Pol III transcription machinery. When the *Alu* B box was mutationally inactivated (by substituting CG for the invariant TC dinucleotide of the B box consensus sequence GWTCRAnnC), a dramatic reduction in *Alu* transcription efficiency was observed, thus confirming the essential character of this element for *Alu* transcription (30).

Upstream flanking sequences have previously been shown to influence transcription efficiency of *Alu* and other SINEs both *in vitro* and in transfected cells. In particular, upstream deletion mutants of an individual *Alu* element displayed reduced transcription efficiency, possibly due to the loss of interactions with sequence-specific transcription factor(s) (28). In another study, upstream sequences already known to stimulate transcription of Pol III-transcribed genes (such as vault or U6 RNA genes) were shown to stimulate SINE transcription in chimeric constructs (29). To explore more extensively the role of upstream regions in *Alu* transcription, we constructed 5' deletion mutants of the 9 isolated *Alus* and compared their *in vitro* transcriptional activity with the one of wild type constructs.

In each case, the natural upstream sequence up to position -12 (or -15 for *Alu*Sx_chr10) was replaced by vector sequence, and care was taken to have each wt-deleted *Alu* pair inserted into plasmid vector with the same orientation, to minimize differences in transcription due do different vector sequence contexts. As shown in Figure 18, as a general trend, upstream sequence deletion negatively affected transcription; however, the extent of transcription inactivation varied markedly among the different *Alus*. Transcription of upstream deleted *Alus* was reduced by 4 to 5 fold in the case of *Alu*Sx_chr1 and *Alu*Sx_chr10 and *Alu*Sp_chr17 (cf. lanes 7, 25 and 27 with lanes 8, 26 and 28, respectively), while it was not appreciably affected in *Alu*Y_chr10-a (lanes 19 and 20) and only moderately reduced (~1.5 fold) in the case of *Alu*Sq2_chr1 and *Alu*Sx1_chr3 (cf. lanes 5 and 9 with 6 and 10, respectively). Overall the data consolidate the notion that the nature of the upstream region may strongly influence *Alu* transcription; however, they do not reveal any obvious correlation between upstream sequence dependency and *in vivo* expression profiles.

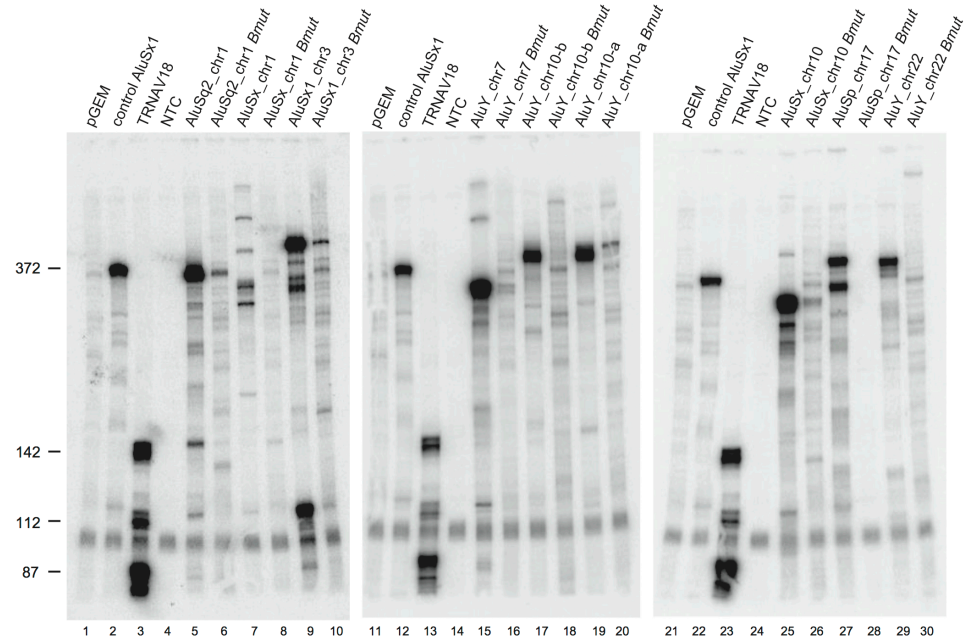


Figure 17. **In vitro transcription analysis of wild type and B box-mutated** *Alu* **loci.** *In vitro* transcription reactions were performed in HeLa nuclear extract using 0.5 mg of the indicated *Alu* templates (lanes 5–10, 15-20, 25-30). A previously characterized *Alu* producing a 372-nt RNA (lanes 2, 12, 22) and a human $tRNA^{Val}$ gene producing a known transcript pattern due to heterogeneous transcription termination (lanes 3, 13, 23) (31) were used as positive controls for *in vitro* transcription and, at the same time, as a source of RNA size markers. Negative control reactions contained either empty pGEM®-T Easy vector (lanes 1, 11, 21) or no template DNA (no-template control (*NTC*), lanes 4, 14, 24). For each *Alu*, both the wild type and a B box-mutated (*Bmut*) version were tested.

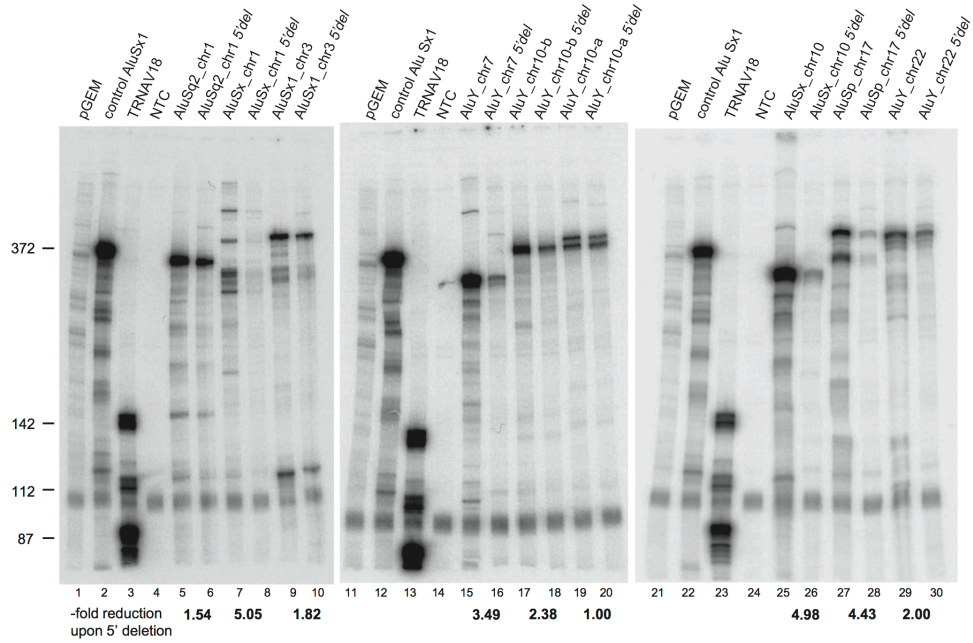


Figure 18. **In vitro transcription analysis of upstream deleted *Alu* loci.** *In vitro* transcription reactions were performed in HeLa nuclear extract using 0.5 mg of the indicated *Alu* templates (lanes 5–10, 15-20, 25-30). A previously characterized *Alu* producing a 372-nt RNA (lanes 2, 12, 22) and a human $tRNA^{Val}$ gene producing a known transcript pattern due to heterogeneous transcription termination (lanes 3, 13, 23) (31) were used as positive controls for *in vitro* transcription and, at the same time, as a source of RNA size markers. Negative control reactions contained either empty pGEM®-T Easy vector (lanes 1, 11, 21) or no template DNA (no-template control (*NTC*), lanes 4, 14, 24). For each *Alu*, both the wild type and a mutant version lacking most of the native 5'-flanking region (*5'del*) were tested. For each of the nine *Alus* subjected to 5'-flank deletion, the extent of reduction of transcription activity, observed with respect to the corresponding wild type *Alu*, is reported below the lanes corresponding to each wt-mutant pair. The values represent the average of two independent transcription experiments that differed by no more than 20% of the mean.

4.1.8. Association with transcription factors of expression-positive *Alus*

The influence of upstream region on *Alu* transcription might be mediated by transcription factors (TF) specifically interacting with this region. The availability of ChIP-seq datasets for several transcription factors (TF) within ENCODE prompted us to assess whether the *Alus* identified as expressed through RNA-seq data analysis tend to be associated with one or more Pol II TF, in addition to the known components of the Pol III machinery. The results of this analysis are

reported in detail in Supplementary Table S3. Since the different cell lines selected for our study have been subjected to ChIP-seq analyses for a highly variable number of TFs (https://genome.ucsc.edu/ENCODE/dataMatrix/encodeChipMatrixHuman.html), a high variability of TF association was observed among them, both in terms of total number of TF-bound *Alus* (ranging from 17 to 79) and in terms of the number of TFs associated to each *Alu*. As a general trend, intergenic/antisense *Alus* tend to be strongly enriched for the presence of TFs associated with their upstream region, with respect to gene-hosted *Alus*. Apart from the components of the Pol III transcription machinery (including TBP), the transcription proteins most frequently associated to expressed *Alus* in most cell lines were the transcription regulator and genome organizer CCCTC-binding factor (CTCF), RNA polymerase II (detected through its largest subunit Rpb1) and the Pol II transcription factor JunD. All of these proteins have previously been shown to colocalize with active Pol III-transcribed loci (especially tRNA genes), where CTCF might contribute to the increasingly recognized function of these loci in nuclear organization and insulation (104-107). Their association to expression-positive *Alus* thus strengthens the notion that the expression-positive *Alu* loci identified in this study resemble the other Pol III-transcribed genes not only for their transcription properties but also for their extra-transcriptional function in genome organization. Furthermore, our analysis revealed a clear cell line-specific association of expression-positive *Alus* with CCAAT/Enhancer Binding Protein β (CEBPB), which was one of the two most frequently matching TFs in HeLa, HepG2 and K562 cells (see Supplementary Table S3; P-values for enrichment of CEBPB at expression-positive *Alu* loci with respect to total *Alus* were lower than 10^{-5}, 10^{-10} and 10^{-6} for HeLa, HepG2 and K562 cells, respectively). This protein was not previously reported to be enriched at other Pol III-transcribed genes; it might thus represent a novel, *Alu*-specific TF facilitating *Alu* transcription.

ChIP-seq analyses have provided recently a considerable wealth of information on histone modification marks at Pol III-transcribed genes, revealing a broad similarity between epigenetic marks typical of active Pol II- and Pol III-transcribed genes, together with a few possibly significant differences (104). The search for histone modification profiles typical of expressed *Alu* loci on the basis of ENCODE ChIP-seq data, that we performed by focusing on purely intergenic expressed *Alus*, did

not produce easily interpretable results, possibly because of the too low number of analysed loci (data not shown). From the data in Supplementary Table S3, however, we noticed that in HepG2, HeLa and K562 cells the P300 acetyltransferase (EP300) is among the top ten TFs associated to expression-positive *Alus* (with a P-value $< 10^{-12}$ for enrichment at expression-positive *Alu* loci with respect to total *Alus* in both HeLa and K562 cells) thus suggesting that these *Alus* might be characterized by high levels of histone acetylation, in agreement with the results of a recent study showing an enrichment of H3K27ac and P300 at enhancer-like *Alu* elements (52).

4.2. ENCODE: MIRs

4.2.1. A bioinformatic pipeline for the identification of transcriptionally active MIR loci from RNA-Seq datasets

To leverage the potential of the search algorithm employed above for *Alu* transcriptome profiling, we developed a Python script in order to improve the identification of genuine Pol III SINE transcripts (see Methods and Supplementary Materials). Here we focused on the identification of MIR transcripts by applying the *SINEsFind* bioinformatic pipeline (par. 3.2.2), summarized in Figure 11, to the previously aligned Long RNA-Seq reads from ENCODE.

We considered in our analysis a subset of the whole annotated MIR elements from hg19 UCSC Table browser Repeatmasker track classified according to their genomic location: 1) MIR mapped in intergenic region or within RefSeq, Ensembl and lincRNA genes but in antisense orientation (from here on named "intergenic/antisense") and 2) MIR fully contained within introns of RefSeq, Ensembl and lincRNA genes in sense orientation and not overlapping with any exon.

All the presented results arise from uniquely aligning RNA-seq reads, being the unambiguous identification of expressed MIR elements our main goal.

4.2.2. General features of MIR transcriptomes emerging from ENCODE RNA-seq data analysis

A preliminary survey of the expression coverage of the MIR loci identified as expression-positive showed us both convincing MIR transcription profiles and less clearcut profiles with noise background signals upstream and downstream the MIR element probably deriving from the sequencing of introns or unknown Pol II transcripts. Even the latter cases nevertheless showed expression coverage enrichment in the region of the annotated MIR element opening the possibility of its Pol III transcription, concomitant to Pol II transcription of the hosting gene, or even its intron retention as part of longer Pol II transcripts. Since the noise background signal is more frequently observed in MIR element mapping within introns of Pol II genes in sense orientation, which constitute the 83% of the

expression-positive MIRs found (see Supplementary Table S5), we decided to mainly focus on the expression of intergenic/antisense MIRs, while a few examples of gene-hosted MIRs will be addressed later in the Results section.
For all the expression-positive intergenic/antisense MIRs found we performed a bioinformatic analyses using the Pol3scan program (108) (see Supplementary Materials) on the corresponding expected full-length MIR aimed at evaluating the presence and conservation of the Pol III functional promoters A- and B- boxes (see Supplementary Materials).

The full list of MIRs identified as expressed by our search strategy is reported in Supplementary Table S5. As summarized in Table 5 each of the cell lines expressed a limited number of MIR elements (ranging from 135 in the case of HeLa cells to 310 in the case of HUVEC cells). Of the whole set of 1097 expression-positive MIR loci a small percentage (17%) were intergenic/antisense, which is in contrast with the global distribution of all the annotated MIRs used in the pipeline where the percentage of the intergenic/antisense MIRs is much higher (66%) (see Supplementary Table S5). Of these a significant percentage (~38%) actually mapped in antisense orientation to annotated Pol II-transcribed genes, in agreement with the distribution of the whole dataset used (42%). However it is worth noting that the fraction of intergenic/antisense expression-positive *Alus* mapping in antisense orientation to the hosting genes is only 22%, thus suggesting intronic MIRs could have a more specific role correlated with Pol II genes regulation. A small fraction (~17%) of all the expression-positive MIRs were found to be expressed in more than one cell line (see Supplementary table S5), while this percentage raise down to 11% considering only the intergenic/antisense ones (Table 5). This suggests a marked cell line specificity in MIR expression (even though few MIRs seem to be ubiquitously expressed and will be further analyzed later on) greater than the one showed in *Alu* expression where the fraction of expression-positive *Alus* found to be expressed in more than one cell line was greater (~24%).
As summarized in Table 6, no significant under or overrepresentation of any particular MIR subfamily within the set of expressed MIR was observed, except for MIR3 that appeared to be slightly underrepresented when considering both the total genomic (intergenic/antisense and intronic sense) and the intergenic/antisense sets of expression-positive MIR (see Supplementary Table S5).

Table 5. Statistic of expression-positive MIR elements in selected cell lines

Cell line	Total MIRs[1]	Intergenic/antisense	Intergenic/antisense shared[2]	Antisense to introns[3]
GM12878	168	32	5	20
H1-hESC	228	39	10	14
HeLa-S3	135	25	7	5
HepG2	247	42	6	14
HUVEC	310	35	3	4
K562	146	37	10	14
NHEK	147	33	12	16
ALL[4]	1097	188	20	71

[1] For each cell line, the column reports the number of MIRs considered as autonomously expressed in both ENCODE RNA-seq replicates.

[2] For each cell line, the column reports the number of intergenic MIRs that are also expressed in one or more different cell lines.

[3] Reported in this column are the numbers of intergenic MIRs mapping with an antisense orientation to introns of both protein-coding and lncRNA genes.

[4] The numbers in this raw refer to individual MIRs expressed in one or more cell lines.

Table 6. Subfamily distribution of expression-positive MIRs

MIR subfamily	Total genomic[1]	Expressed genomic[1]	Total intergenic/ antisense[1]	Expressed intergenic/ antisense[1]
MIR	174175 (30%)	352 (32%)	116136 (30%)	71 (38%)
MIRb	223577 (38%)	481 (44%)	148950 (38%)	84 (45%)
MIRc	102688 (17%)	162 (15%)	68069 (17%)	23 (12%)
MIR3	90185 (15%)	102 (9%)	59209 (15%)	10 (5%)

[1] Reported are the absolute copy numbers and (in parentheses) the percentages of MIRs of each sub-family considered relative to (from left to right): the total set of genomic MIRs ("Total genomic"); the set of MIRs found to be expression-positive in one or more cell line ("Expressed genomic"); the total set of intergenic/antisense MIRs; the set of intergenic/antisense MIRs found to be expression-positive in one or more cell lines;

4.2.3. Survey of expressed MIRs according to location and base-resolution expression profile

The ~ $6x10^5$ annotated MIRs are not all complete in sequence, implying that many of them represent only a portion of the canonical full-length MIR element. Among expression-positive MIRs, we found both complete and incomplete elements and, correspondingly, four main types of base-resolution expression profiles: 1) full-length or almost full-length MIRs (Figure 19A-B), covered by sequence reads along all their extension; 2-3) incomplete MIRs representing either the left or the right portion of the canonical full-length MIR but whose transcript coverage tends to correspond to the one of a fully transcribed canonical MIR More precisely, transcript coverage tends to extend into the downstream MIR-unrelated region for incomplete MIRs lacking the 3' moiety (Figure 19C), while it tends to start in an upstream MIR-unrelated region, possessing functional A and B-boxes, for incomplete MIRs lacking the 5' moiety (Fig 19E). 4) In a fourth type of profile, an incomplete MIR lacking the 3' moiety (thus containing A and B boxes) produces transcripts that do not extend outside of the MIR sequence (Figure 19D).

The MIR reported in figure 19A (chr14:34206132-34206363 MIR_dup717) was scored as expression-positive in Gm12878 and K562 cells. It was also expressed at very low levels, below the threshold for a positive scoring, in H1-hESC, HeLa-S3, HepG2 and NHEK. This representative MIR is full-length and located within intron 6/7 (depending on which transcript isoform is considered) of NPAS3 gene, in antisense orientation. Paradigmatically, it shows an expression profile that completely covers the annotated element with uniquely mapped sequence reads. A- and B-boxes are conserved at 13 and 50 bp downstream of the TSS, respectively. Unexpectedly the annotated MIR contains a strong termination signal (T_4) in its sequence 104 nt from the beginning of the element, which is clearly skipped by the Pol III machinery. We investigated the possibility that the genome sequences of the cell lines from which RNA-Seq reads arise were mutated in correspondence of these strong terminator sequences. To this end, we reconstructed the corresponding consensus genome sequences from RNA sequence reads (see Supplementary Materials). We did not find any sequence variant, therefore indicating that Pol III machinery truly skip these terminator signals. Another strong terminator of 4Ts is

present in the downstream moiety of the MIR, leading to a transcript of ~220 nt. A strong support for the authenticity of this Pol III-derived MIR RNA comes from the binding at this locus of RPC155 in K562 cells.

Reported in Figure 19C is the transcript coverage profile of an incomplete MIR of 159 bp (chr16:22309780-22309939 MIR_dup2285), which turned out to be expressed in all 7 cell lines. This element aligns to the left portion of the canonical MIR sequence, reported in green in the figure panel. Transcription of this element appears to continue downstream in a MIR-unrelated region for ~200 nt before encountering a non-canonical termination signal (T_3CT). This previously characterized MIR (42) is located in antisense orientation in the first intron of the POLR3E gene, encoding a subunit (POLR3E/RPC5) of human RNA polymerase III. Its nature of Pol III transcription unit is supported by the presence of canonical A- and B-boxes as well as by association with components of the Pol III machinery in HeLa-S3 and K562 cell lines (see Figure 19C). Here, again, the annotated MIR contains a strong termination signal (T_4) in its sequence, 121 nt downstream the 5' end of the annotated element, and another one 80 nt downstream the 3' end of the annotated element at which, however, expression coverage signal tend to decrease temporarily but slightly increase again after it until the non canonical terminator signal is encountered. The skipping of the first strong terminator by Pol III is supported by the expression coverage levels before and after it which are almost the same strongly suggesting that they arise from the same transcript. Indeed the expression coverage decrease, which accidentally abuts precisely the strong terminator, is due to RNA sequencing specifications and cDNA size selection during library preparation, as previously explained. Investigating the possibility of a mutation in the these strong terminators sequences in the DNA of these cells (see above), we did not find any sequence variant, thus confirming the ability of Pol III to skip strong terminator signals within certain sequence contexts (31). This finding is partially confirmed by *in vitro* transcription analysis performed using HeLa nuclear extract (see below). Indeed the size of the most abundant transcript correspond to the one ending at the first strong Pol III terminator sequence while weaker signals correspond to transcripts ending at the other terminators found in the cloned sequence. Thus while the strong Pol III terminator seems to be clearly skipped in vivo, it is only partially skipped in vitro maybe due to the absence of unspecified chromatin features.

Intriguingly an antisense-oriented MIR with the same location and similar sequence is also present in the mouse genome, where it has also shown to be transcriptionally active (44). To determine whether Pol III transcription of this MIR affected in some way Pol II transcription of the host gene, the previous study examined Pol II occupancy over the POLR3E gene, as well as H3K4me3 and H3K36me3 histone marks on this region, and found an increased accumulation of H3K36me3 in a 3' direction within the body of the gene and an unusual second point accumulation of RNA Pol II, in addition to the expected one on POLR3E promoter, which abuts precisely on RNA Pol III occupancy peak which reflect MIR transcription on the opposite strand. Thus, active antisense Pol III transcription of the MIR apparently creates a barrier for Pol II, which, as a result, slows down and accumulates just at the downstream border of the MIR.
These observations, along with the fact that this MIR is conserved in human and mouse, led the authors to hypothesize a functional role of the MIR in POLR3E gene regulation. We therefore asked if this second RNA Pol II accumulation point also occurs in the ENCODE cell lines, and we found Pol II peaks for all the 7 cell types (see Supplementary Table S7) thus strengthening the hypothesis that this transcriptionally active MIR plays a regulatory role in Pol II transcription of POLR3E. Figure 20 shows this second accumulation of Pol II for three of the seven cell lines for which are also available (and shown) signals of Pol III.

Figure 19C shows a MIR, found as expression-positive in H1-hESC, K562 and NHEK cell lines, whose transcription initiates in an upstream MIR-unrelated region (chr14:89445565-89445634 MIRc_dup2189) and ends ~45 nt downstream of the annotated element in correspondence of a strong termination signal (T_5). Inspecting the MIR-unrelated upstream region we could find canonical A- and B- boxes supporting the possibility that this MIR element lacking Pol III promoters in its sequence is transcribed from an upstream unrelated MIR region providing functional control elements. Here again we noted a potentially strong termination signal (T_3AT_4) at the beginning of the annotated element, but the binding of Pol III throughout this region in K562 cells, together with the transcript coverage profile, strongly support the existence and Pol III-dependence of this MIR transcript. That Pol III can at least partially read through the internal T_3AT_4 element is also supported by *in vitro* transcription data (see below).

In Figure 19D we report a MIR transcript (chr17:17863550-17863651 MIRb_dup1281), found to be expressed at low levels in in H1-hESC cell line, originating from an incomplete MIR element antisense to TOM1L2. Inspecting the coverage profile in other cell lines, we could found signals of much lower expression in at least one replicate of each of the analyzed cell lines. The source element of this transcript corresponds to a MIRb left fragment carrying in its sequence functional A- and B- boxes. The transcript coverage precisely spans the whole length of the element until a strong terminator (T_{10}) right at the end of it, leading to a transcript of ~100 nt.

As mentioned above, inspecting coverage profiles of the expression-positive MIRs we found that some of them were also present in other cell lines but with coverage levels under the background signal in one of the two replicates, while the other expressed it even with over 2 order of magnitude, not due to the different sequencing depth. For this reason they did not pass the filter of our bioinformatic pipeline which required the presence of the MIR transcripts in both replicates. This could mean that MIR expression is not only cell specific but also dependent on other various factor (e.g. growth conditions).

Finally we asked if there was a correlation between the expression of MIRs hosted in Pol II genes in antisense orientation and the expression of host genes themselves. To this end, we compared normalized read counts of the MIRs in each replicate with those of the hosting Pol II genes (Se Supplementary Materials). We found a direct correlation (using Pearson coefficient) between MIR and host gene expression in the case of NRXN1 (0.998), HSPG2 (0.98), C9orf91 (0.93), KCNJ6 (0.86), HIVEP3 (0.8), while RAPH1 (-0.74) was inversely correlated (data not shown)

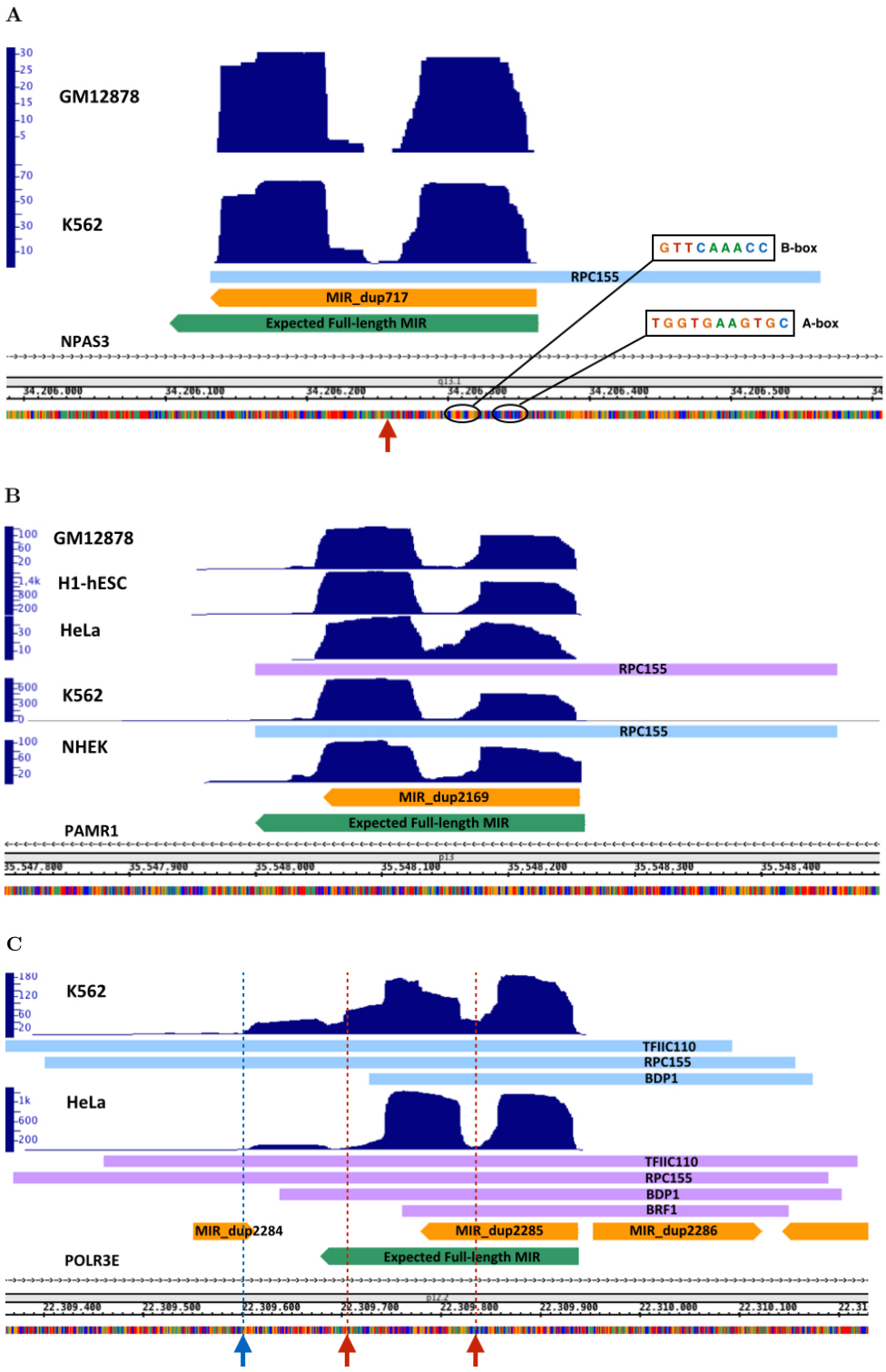


Figure 19. See below for description

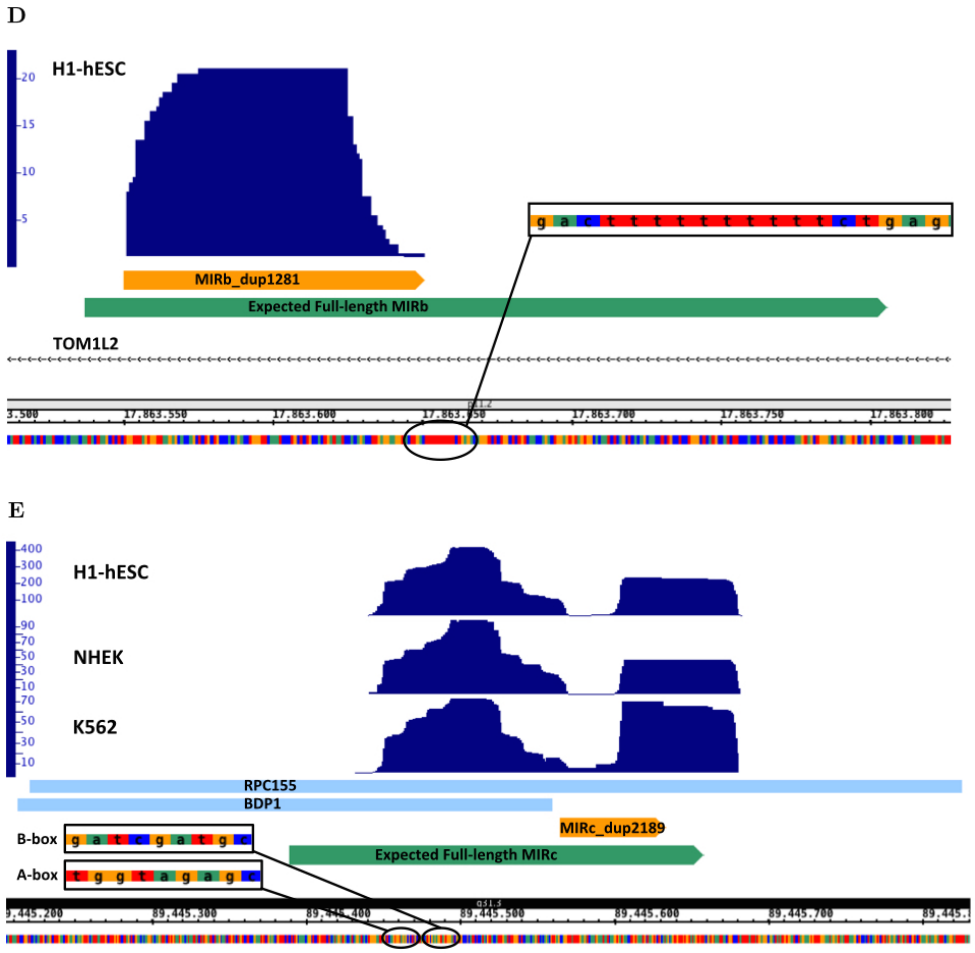


Figure 19. Base-resolution expression profiles for five representative MIRs. See text for descriptions. Bars in light blue and magenta represent component of the Pol III transcription machinery bound at the corresponding loci; orange arrowed bars represent the annotated MIR elements while the green arrowed bars represent the corresponding expected full-length MIR elements which do not correspond to annotated elements and are reported merely to locate the alignment position of the annotated MIRs (orange) inside the corresponding consensus sequence. Red and blue arrows show respectively the positions of the strong and non-canonical Pol III terminators. A) MIR_dup717 chr14:34206132-34206363; B) MIR_dup2691 chr11:35548054-35548257 which reside inside an intron of the PAMR1 gene in sense orientation; C) MIR_dup2285 chr16:22309780-22309939; D) MIRb_dup1281 chr17:17863550-17863651; E) MIRc_dup2189 chr14:89445565-89445634

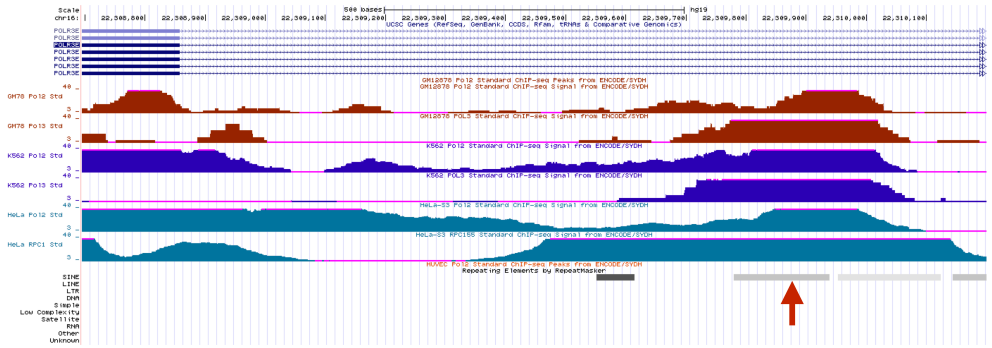


Figure 20. Pol II accumulation signals. Shown are the signals of Pol II (POLR2A) and Pol III (POLR3G) in GM12878 (red) and K562 (blue). The .cyan profiles represent signal of Pol II (POLR2A) and Pol III (RPC155) in HeLa cells. The red arrow points to the expression-positive MIR inside the first intron of POLR3E gene in antisense orientation. (Figure 19C)

4.2.4. Association of the Pol III machinery to expression-positive MIRs

Before the use of RNA-Seq data to help identify expression-positive SINEs, genome-wide studies based on ChIP-Seq approaches were used with the aim of producing inventories of loci inferred to be transcribed by Pol III from their association with one or more component of the Pol III machinery (42, 44). Such studies revealed a limited number of Pol III-associated SINE elements, and almost the totality of them were *Alus*. The availability of genome-wide ChIP-Seq data from ENCODE/Stanford/Yale/USC/Harvard (SYDH) for key components of Pol III transcription machinery (Bdp1, Brf1/2, Rpc155 and TFIIIC110) allowed us to re-address this issue, by investigating whether a significant enrichment of these TFs could be found in the expression-positive intergenic/antisense MIRs of HeLa and K562 cells. Supplementary Table S6 lists the expression positive MIRs found to be associated to Pol III components in Hela and K562 cells. In HeLa cells only 3 intergenic/antisense expression-positive MIRs, together with 1 MIR hosted in the PAMR1 gene in sense orientation, were found associated to 1 or more Pol III TFs of the 15 intergenic/antisense and of the 120 intron-hosted expression-positive MIRs. The only association found to be statistically significant, with a P-value of 1.5×10^{-6} (calculated using Fisher's exact test), was the Bdp1 component when considering the fraction of the intergenic/antisense expressed MIRs against the total intergenic/antisense annotated ones. When K562 cells were considered, a higher percentage (27%) of intergenic/antisense expression-positive MIRs where found

bound by one or more Pol III TFs, while the percentage drop down to 3% when considering those fully contained inside introns of Pol II genes in sense orientation (data not shown). In K562 cells we found significant enrichment for Bdp1, TFIIIC110 and RPC155 (P-values $5.4x10^{-10}$, $2.4x10^{-6}$ and $<2.2x10^{-16}$ respectively) in intergenic/antisense expression-positive MIRs. It is interesting to compare these results with those obtained in the case of *Alus* for which even the K562 cells show increased significativity and number of *Alu* loci bound to components of the Pol III machinery complex versus HeLa cells, thus suggesting a more permissive environment for Pol III factor association.
Interestingly the MIR antisense to POLR3E, expressed in all the 7 cell lines, is the one bound by the highest number of Pol III TFs in both HeLa and K562 cells (4 and 3 respectively) thus further confirming its genuine character of Pol III transcription unit.

4.2.5. Association with TFs of expression-positive MIRs

In order to assess whether the upstream region of MIR elements could influence their transcription by TFs specifically interacting with it, we took advantage of the availability of ChIP-Seq data for several TFs within ENCODE. We asked if intergenic/antisense MIRs identified as expressed through RNA-Seq data analysis tend to be associated with one or more Pol II TFs, in addition to the known components of the Pol III machinery. The results of this analysis are reported in detail in Supplementary Table S7. A high variability of TFs association was observed among the 7 cell lines, both in terms of total number of TF-bound MIRs (ranging from 3 to 22) and in terms of the number of TFs associated to each MIR. Among the transcription proteins most significantly associated with expression-positive intergenic/antisense MIRs we found RNA polymerase II (POLR2A), TBP, MAZ, YY1 and PML. Intriguingly the YY1-binding site is known to function as a component of the LINE-1 core promoter to direct accurate transcription initiation (109) and since MIR are thought to have arisen from the fusion of a tRNA with the 3' end of a LINE, the enrichment of the YY1 TF on expression-positive MIR suggests a possible role in MIR transcription initiation that should be further investigated.

HeLa, HUVEC and NHEK cell lines are those with the lowest number of TFs being subjected to ChIP-seq and therefore have very few number of intergenic/antisense expression-positive MIRs bound by TFs and with no significant enrichment.

4.2.6. Expression-positive MIRs and chromatin states

Stimulated by the results of a recent study that revealed *MIR* elements to be highly concentrated in enhancers of K562 and HeLa cell lines (41), we investigated whether our intergenic/antisense expression-positive *MIRs* found in the same cell lines were among those found by this study. We intersected the coordinates of our (K562 and HeLa) expression positive MIRs with those of the MIR elements found to be enriched in enhancer state of the chromatin (lifted over to GRCh37/hg19), but we did not find any overlap. Because of this result, we decided to check if, independently from expression, the annotated intergenic/antisense MIRs used in our bioinformatics pipeline where overrepresented in enhancer states of the chromatin (weak and strong) versus other states using ENCODE Chromatin State Segmentation by Hidden Markov Model (HMM) (110) for each of the seven cell lines (excluding HeLa-S3 for which data are not available). Again, we could not find any enrichment (Supplementary Materials). Though we tested for enrichment in the enhancer states of our intergenic/antisense expression positive MIRs performing a Fisher's exact test against all the other annotated intergenic/antisense MIRs used in the pipeline. We found statistical enrichment only in Gm12878 and HepG2 cell lines in strong enhancer state (P-values 1.876e-^{05} and 2.133e-^{05} respectively) (Supplementary Table S8).

4.2.7. In vitro transcription analysis of expressed MIR elements

Our bioinformatic pipeline permitted us to detect *in vivo* expression of individual MIR elements that could be transcribed by the Pol III machinery. To confirm Pol III transcription and make a precise promoter characterization of these transcriptional units, we conducted their *in vitro* transcription in HeLa nuclear extract. We focused our attention on a small subset of MIR loci, which are expressed in at least three cell types. These loci are listed in Table 2. One of them,

MIR_dup2285 (chr16:22309780-22309939), appeared to be expressed in all the seven investigated ENCODE cell lines, and was also found to be associated with at least 3 components of the Pol III machinery in both HeLa-S3 and K562 cells (see Supplementary Table S6). This transcription unit possesses a non-canonical terminator (T_3CT) ~200 bp after the annotated element and has also a 2 strong early terminators, along with other 1 non canonical terminator signals, that seems to be skipped by the Pol III machinery. The first strong terminator (T_4) is located inside the annotated element at 121 nt from the start coordinate, while the other one (T_4), located at 240 nt downstream the 5' end, is outside the annotated element but inside the corresponding expected full-length MIR. As shown in Figure 21, all the 4 transcripts, corresponding to the 4 termination signals, are identified during *in vitro* transcription, supporting the hypothesis of strong and non-canonical terminator skipping by Pol III. However, as discussed above, it is worth noting that the most abundant signal arise from the transcript ending at the first strong Pol III terminator, thus suggesting a limited ability of the Pol III machinery complex to skip its terminators in vitro. A fifth transcript is also identified corresponding to a transcript ending to the first strong terminator but using an alternative A-box (see Table 2).

Another locus subjected to in vitro transcription analysis, MIR_dup3493 (chr1: 34943459-34943727), was found to be expressed in four cells lines (GM12878, H1-hESC, K562 and NHEK) and has perfect A- and B- boxes, according to the *consensus* ones (27). The remaining two in vitro tested loci (MIRb_dup5848 chr2:71762977-71763215 and MIRc_dup2189 chr14:89445565-89445634) were found to be expressed in three cell types (H1-hESC, HepG2, NHEK and H1-hESC, K562, NHEK respectively) and, based on ENCODE ChIP-seq data, only MIRc_dup2189 is associated with two components of the Pol III machinery (RPC155 and BDP1).

For each of the four expression-positive MIRs, the corresponding expected full-length sequences were PCR-amplified from human genomic DNA, cloned into pGEM-T-Easy vector and their ability to support efficient *in vitro* transcription was tested using HeLa cell nuclear extract. To be sure that the observed transcripts were produced by Pol III transcriptional machinery, reactions were conducted in the presence of α-amanitin at a concentration (2 µg/ml) known to completely inhibit RNA polymerase II activity. Transcription reactions were also planned in parallel

with two mutant versions of each MIR element: one in which the B-box internal promoter element was inactivated by site-specific mutagenesis, and the other one in which the upstream flanking region was deleted. The results of *in vitro* transcription analysis are shown in Figure 21.

Control transcription reactions were set with empty pGEM-T-Easy plasmid (lanes 1 and 10) and the same vector carrying either a previously characterized *Alu* (*Alu*Sq2 chr1:61523296–61523586, see Figure 17, lane 5) (lane 2 and 11) and a tRNAVal (AAC) gene (TRNAV18, chr6) (lane 3 and 12) whose transcription produces three different primary transcripts (of 87, 112 and 142 nt) because of heterogeneous termination at one of three consecutive termination signals (31). Each of the tested MIR elements produced a distinct pattern of transcription, in which the sizes of the most abundant transcripts agreed with those predicted on the basis of sequence inspection of the Pol III termination signals, either canonical (a run of at least four Ts) or non-canonical, both internal and in the 3'-flanking region (see Table 2)

Interestingly, transcription efficiencies of all MIRs were roughly comparable, indicating that their different tendency to be transcribed in cultured cells is not due to differences in *cis*-acting elements recognized by the basal Pol III transcription machinery. Indeed, when the MIR B-box was mutationally inactivated (by substituting CG for the invariant TC dinucleotide of the B box consensus sequence GWTCRAnnC), a dramatic reduction in MIR transcription efficiency was observed in each case, thus confirming the importance of this element for MIR transcription, as previously detected for *Alu* transcription (see above)

Transcription of B-box mutants was reduced by 4.4 to 5.2-fold in the case of MIR_dup3493 and MIRb_dup5848 (cf. lanes 7 and 13 with lanes 9 and 15, respectively), and reduced by 2.4- to 3.8-fold in the case of MIR_dup2285 and MIRc_dup2189 (cf. lanes 4 and 16 with 6 and 18 respectively).

These data demonstrate that MIRs are efficiently transcribed by the Pol III transcription machinery and confirm a key role for B box recognition by TFIIIC, even though the appreciable levels of residual transcription observed with B box-mutated MIRs suggest the possibility that MIR promoter strength does not rely entirely on this element. .

The transcription efficiency of *Alus* and other SINEs was also shown to be influenced by 5'-upstream region in both *in vitro* and in transfected cells, and in particular, upstream deletion mutants of an individual *Alu* element displayed

reduced transcription efficiency, possibly due to the loss of interactions with sequence-specific TFs (28). According to these findings, the above reported in vitro transcription analysis of *Alus* data consolidate the notion that the nature of the upstream region may strongly influence *Alu* transcription and To understand if upstream regions could also have an effect on MIR transcription, we compared the *in vitro* transcriptional activity of the four isolated MIRs with that of the corresponding 5'-deletion mutants. As shown in Figure 21, upstream sequence deletion negatively affected transcription to different extents for the different MIRs. Transcription of upstream deleted MIRs was reduced by 1.8- to 2.5-fold in the case of MIR_dup2285, MIR_dup3493 and MIRb_dup5848 (cf. lanes 4, 7 and 13 with lanes 6, 9 and 15, respectively), and only moderately reduced (~1.35-fold) in the case of MIRc_dup2189 (cf. lanes 16 with 18) These data reveal that MIR transcription might be influenced by the upstream region as observed for *Alu* transcription.

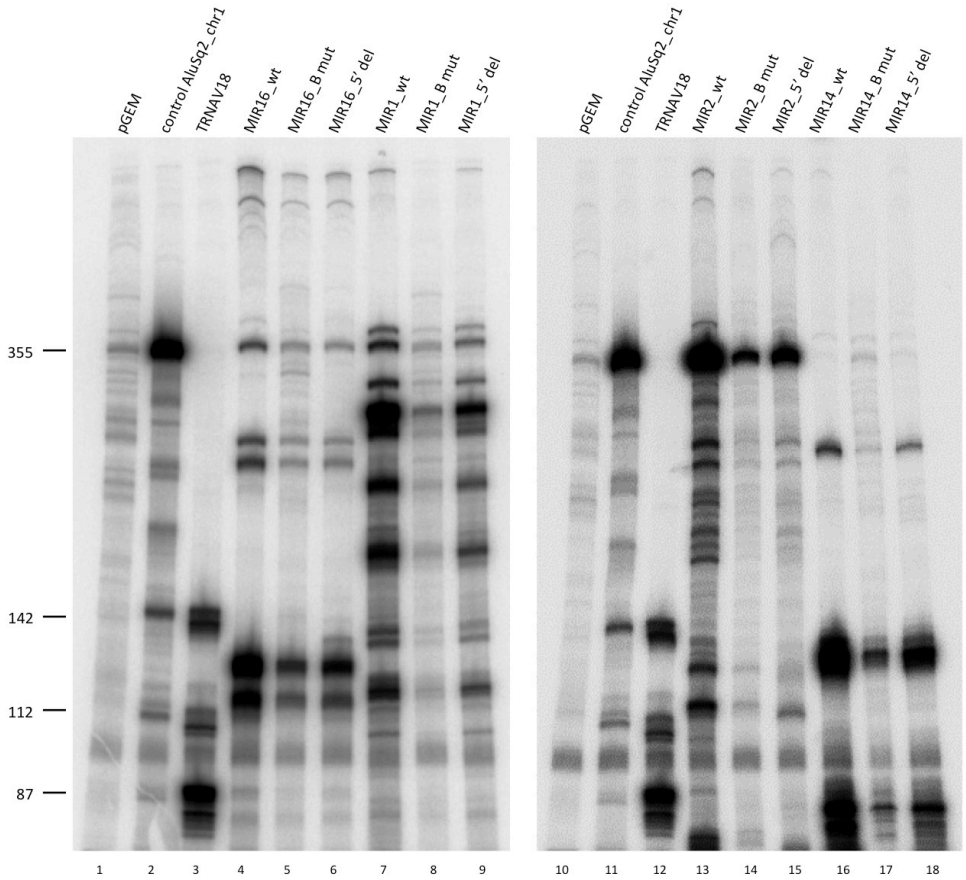


Figure 21. In vitro transcription for selected expression-positive MIRs. *In vitro* transcription reactions were performed in HeLa nuclear extract using 0.5 mg of the indicated MIR templates (lanes 4–9, 13-18,). A previously characterized *Alu* producing a 355-nt RNA (lanes 2, 11) and a human $tRNA^{Val}$ gene producing a known transcript pattern due to heterogeneous transcription termination (lanes 3, 12) (31) were cloned into pGEM®-T Easy vector used as positive controls for *in vitro* transcription and, at the same time, as a source of RNA size markers. Negative control reactions contained empty pGEM®-T Easy vector (lanes 1, 10). For each MIR, both the wild type, B box-mutated (*Bmut*) and 5'-flanking region (*5'del*) version were tested.

4.3. *Alu* expression profiling in *dl*1500 Ad5-infected IMR90 cells

4.3.1. General features of *Alu* transcriptomes

The availability of RNA-seq and ChIP-seq data of an oncogenic in vitro model (see Material and Methods), along with the proven effectiveness of our bioinformatics pipeline in identifying expression-positive SINE loci, prompted us to investigate Pol III-derived *Alu* RNAs deregulation.

By applying the improved *SINEsFind* bioinformatics pipeline, we analyzed RNA-seq data representing the transcriptome of IMR90 primary fibroblasts that had been either infected with *dl*1500 Ad5, or mock-infected, for 6 or 24 hours. As a general mark, it is worth noting that since we did not have any biological replicates for RNA-seq, data reported for *Alus* and differential gene expression (see below) couldn't be tested for significativity. However, the availability of samples infected whose infection has been performed at different time, allowed us to consider them as a sort of biological replicates. In mock-infected cells, 150-200 *Alu* elements were identified as significant sources of transcripts (Supplementary Table S9). These *Alus* were classified as expressed because each of them had a peak coverage expression value over 5 (reads) (see Supplementary Materials for details). As mock-infected cells were subjected to RNA-seq analysis at two different growth stages (6 hr and 24 hr after the mock treatment) we made sure to verify the degree of overlapping of the subset of expressed *Alus* in the two samples. The number of expression-positive *Alus* in 6-hr mock-infected cells was ~20% greater than in 24-hr mock-infected cells. We found that almost half of *Alu* elements detected as expressed after a 6-hr growth post-infection (p.i.), according to the above criteria, were also found to be expressed after 24 hours of post-infection growth (Supplementary Table S9). It is interesting to note that the novel *Alu*Ya5-derived chimeric transcript found within the ENCODE dataset is also clearly detected in this study even using the "unique" alignment strategy because of the longer RNA-seq reads which confirm the previously identified genic locus as the source of the transcript, among all the nearly identical loci in the clusters of *snaR* genes.

4.3.2. Small e1a-dependent activation of *Alu* loci

Small e1a has previously been shown to induce a complex response in IMR90 cells in terms of time-dependent alteration of the protein-coding transcriptome (111). Upon *dl*1500 Ad5 infection, a remarkable time-dependent increase in the global level of expression of intergenic/antisense *Alu* loci was observed. At 6 and 24 hours p.i., the numbers of read counts mapping to *Alus* were increased by 2.5-fold and 7.5-fold, respectively, with respect to the corresponding mock treatments (Supplementary Table S9).

Supplementary Table S9 analytically reports the expression data for the expression-positive *Alus* (peak expression coverage >5) in each sample and also grouped in three datasets ("unique": 1055 *Alus* expressed in at least one of the four samples; "induced": 665 *Alus* whose expression level is higher in dl1500 24 hr sample than in any other sample; "shared": 72 *Alus* expressed in each of the four samples).

Interestingly, of the 665 *Alus* whose expression is induced in dl1500 24h, 195 (~29%) have no expression in the other three samples and represent newly activated transcriptional units, while of the 1055 unique expression-positive *Alus* found, 86 show a decreased expression in dl1500 24h, among which 54 show no expression at all thus being silenced 24h p.i.

The e1a-dependent increase in *Alu* RNA levels can be better appreciated through the graphs reported in Figure 22A, showing the average expression profiles across the body of the expression-positive *Alus* listed in Supplementary Table S9 ("unique" dataset) in the four different samples, plus and minus strand. As expected, the coverage by *Alu* transcript reads spans a region of 300-400 bp; here, a strong and time-dependent induction of expression is observed upon Ad5 e1a expression. The coverage signal upstream of the average expression profile on the plus strand is due to few *Alus* located immediately downstream highly expressed 3'UTRs of protein coding genes in sense orientation and that we decided not to filter out. When analogous profiles were generated using the "induced" dataset in Supplementary Table S4, the increase in expression turned out to be even stronger (Figure 22B).

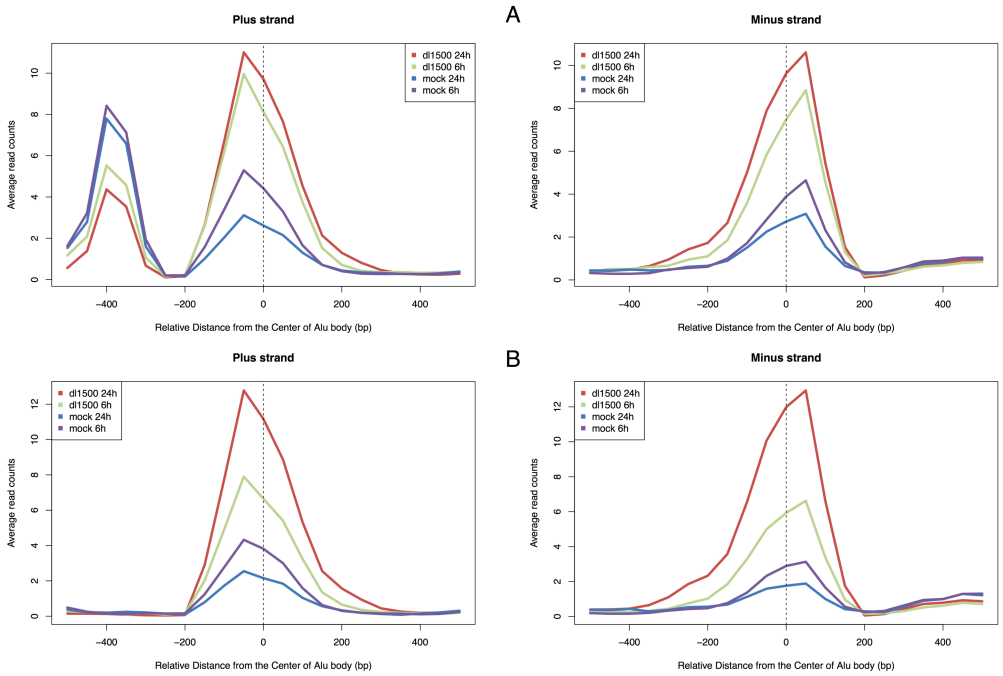


Figure 22. Average expression profiles of A) "unique" and B) "induced" *Alus* datasets across all the samples on plus and minus strands

To start to understand the mechanisms of *Alu* transcriptional activation by e1a, the association of Pol III transcription components with expression-positive *Alus* was investigated through a ChIP-seq approach (see Methods and Supplementary Materials). We found that 138 out of the 774 expression-positive *Alus* in *dl*1500 cells 24h p.i. were associated to one or more Pol III transcription components (15 to BDP1, 3 to POLIII/RPC39 and 133 to TFIIIC-110), while 30 out of the 158 *Alus* found to be expressed in mock cells 24h p.i. where associated only to TFIIIC-110 among which 26 were still bound to it in dl1500 24 hr infected cells but did not show stronger association (See Supplementary Table S10). A previous study reported that e1a activates TFIIIC complex by selective induction of TFIIIC-110 subunit (112) even though a more recent one argues against the model that Pol III transcription can be effectively modulated through the specific induction of TFIIIC110 (113). However, as discussed below, we did not find increased expression for any TFIIIC subunit gene and the number of annotated intergenic/antisense *Alu*

loci bound by TFIIIC-110 remains almost the same following *dl*1500 infection (984 in mock vs 1044 in *dl*1500, of wich 753 are shared) thus arguing against e1a induction of TFIIIC-110, while a slightly increase in BDP1 and POLIII (RPC39) subunits is observed which associate with expression-positive *Alus* only in *dl*1500 sample. Moreover TFIIIC is known to play role in genome organization as insulator through its binding to A- and B- boxes (114, 115) and it is thus not surprising its higher association to *Alu* elements than other Pol III factors. Nevertheless, as shown in Supplementary Table S10, there is a highly significant enrichment of TFIIIC-110 to expression-positive *Alus* (P-val $<2.2e^{-16}$) both in mock and *dl*1500 infected cells. P-values for all the Pol III TFs have been calculated using Fisher's exact test against the whole dataset of annotated *Alus* used (see Supplementary Materials).

4.3.3. Epigenetic context of e1a-dependent *Alu* activation

E1a has been shown to induce extensive epigenome reorganization in IMR90 cells (84, 111, 116). We thus asked whether specific changes in histone modification profiles accompany *Alu* activation at e1a-responsive *Alu* loci investigating whether our responsive *Alu* loci were enriched in of H3K9ac, H3K27ac and H3K18ac histone marks in infected and control cells, by interrogating the corresponding ChIP-seq datasets (see Materials and Methods). We intersected the coordinates of our expected full-length expression-positive *Alus* found in mock and *dl*1500 cells 24 hr p.i. with the corresponding peaks of each histone modification and performed Fisher's exact test against the intersection of the same peaks with the coordinates of the expected full-length *Alus* in the whole dataset used. P-values are reported in Supplementary Table S11.

In mock infected cells we found significant enrichment for all the histone modifications, being H3K18ac the one with the lowest P-value. In *dl*1500 Ad5 infected cells we still found significant enrichment for H3K18ac and H3K9ac but not for H3K27ac.

We though compared these histone modification in both samples by plotting their average ChIP-seq profiles for all the expression-positive *Alus* found in mock and dl1500 cells 24 hr p.i. counted only once (843 "unique" *Alus*) (see Supplementary

Materials). As shown in Figure 23, no differences were observed between mock and *dl*1500 samples except for histone H3 lysine 27 acetylation which showed a ~6 fold increase in mock sample. This is in contrast with the accepted notion that this histone modification facilitates transcription due to more permissive chromatin accessibility. However this result is concordant with the finding that Adenovirus Small e1a decreased H3K27ac at most other promoters than ac1 genes, including promoters of the other e1a-activated clusters, intergenic regions, and introns, resulting in extensive global H3K27 deacetylation (83).

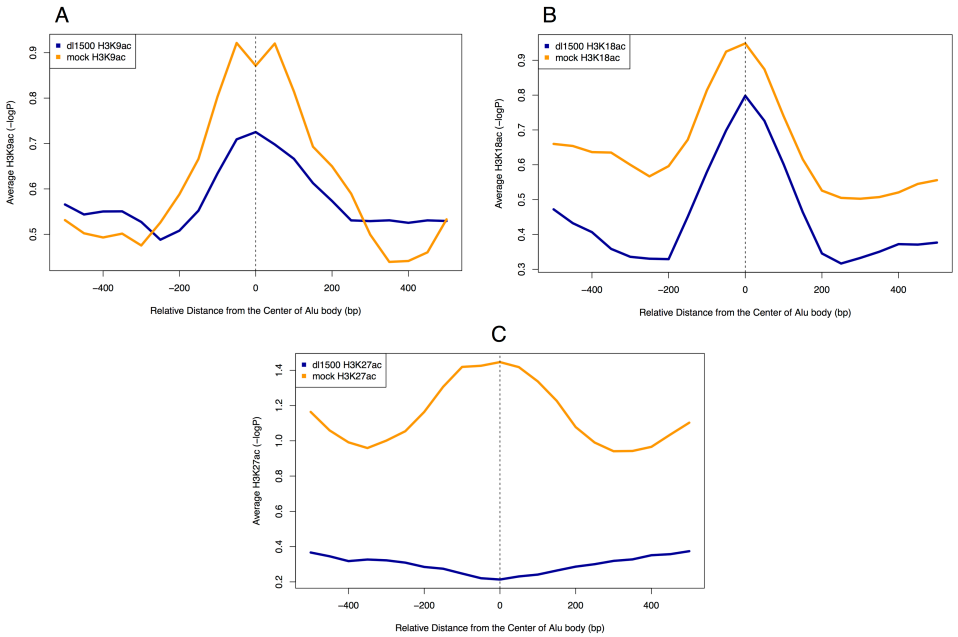


Figure 23. Average histone modifications profiles in mock and dl1500 cells 24 hr p.i.

The analysis of ChIP-seq data for the P300 and RB1 (83), respectively a lysine acetylase and a retinoblastoma protein, revealed a significant enrichment of these proteins to expression-positive *Alus* found by our pipeline in mock and *dl*1500 cells 24 hr p.i. with the exception of P300 in mock sample (Supplementary Table S11). However no variation is detected between mock and *dl*1500 samples when comparing average signal profiles using dataset of "unique" expression-positive *Alus* in 24 hr p.i. samples (data not shown). Interestingly we found colocalization of all

TFIIIC-110 subunits with P300 whose direct interaction with TFIIIC is known to stabilizes binding of TFIIIC to core promoter elements and to be recruited to the promoters of actively transcribed tRNA and U6 snRNA genes in vivo (117)

4.3.4. Distinctive features of responsive *Alu* elements.

The majority of *Alu* elements in the human genome belong to the older *Alu* subfamilies (S+J), while only the young (Y) subfamilies are thought to be retropositionally active (118). Although this observation might lead to consider *Alu* Y as more active transcriptionally, previous studies aiming at identifying expressed *Alu* loci could not evidence any significantly higher expression of younger *Alus* (see par. 4.1.2 and Table 4). We evaluated this point by calculating the *Alu* subfamily distributions of expressed and "responsive" *Alus,*,that is *Alus* whose expression is increased in *dl*1500 cells 24 hr p.i. with respect to all other samples ("induced" dataset, Supplementary Table S9), and comparing them with the subfamily distribution of a collective dataset of 787333 intergenic/antisense *Alus* (this set also included incomplete *Alu* elements, as a few of them were found to be transcribed). As reported in Table 7, a significant variation with respect to the distribution of all the annotated *Alus* used in our pipeline was observed for the *Alu*J and *Alus* subfamilies. In particular, *Alu*J appeared significantly depleted from the induced *Alu* set, while *Alus* (that are largely predominant in both datasets, as expected) are more represented among the induced *Alus* (85%) than among total *Alus* (61%). These results are in contrast with our previous findings and the notion related to the activity of the retroposition process of the different *Alu* subfamilies does not correlate with their transcription/expression levels, which are clearly inverted in this study.

Table 7. Subfamily distribution of "induced" *Alus*

Family	Total intergenic/antisense[1]	Induced[1]	P-val[2]
S	477144 (61%)	566 (85%)	<2.2e-16
J	209701 (27%)	50 (8%)	<2.2e-16
Y	100488 (13%)	49 (7%)	1.176e-05
TOTAL	**787333**	**665**	

[1] Reported are the absolute copy numbers and (in parentheses) the percentages of *Alus* of each sub-family considered relative to the total set of intergenic/antisense *Alus* and the set of intergenic/antisense *Alus* whose expressions are induced in *dl*1500 sample 24 hr p.i.;

[2] P-values has been calculated using Fisher's exact test

4.3.5. E1a-dependent deregulation of other Pol III-transcribed genes and of genes coding for components of the Pol III machinery

Early studies reported an e1a-dependent increase in Pol III transcription of tRNA, 5S rRNA and VA-RNA genes in nuclear extracts of adenovirus-infected HeLa cells (119, 120). Such e1a-dependent activation of gene transcription could also be observed in cultured cells for transfected class III genes, but only marginally for major endogenous cellular class III genes (120). We sought to evaluate the effect of e1a-dependent overexpression on non-*Alu* Pol III-transcribed genes from RNA-seq data. However, since the employed RNA-seq procedure tended to exclude small-sized RNAs such as the 75-80 nt-long tRNAs, and the 5S rRNA also tends to be lost in the rRNA depletion treatment, we could only estimate reliably transcription levels for longer major Pol III transcripts, such as 7SL, 7SK, RNase P and RNase MRP RNAs (Supplementary Materials). In each case we did not observe any e1a-dependent increase in expression but, instead, a slight decrease (data not shown).
It has previously been reported that increased Pol III transcription by e1a correlates with an increase in the expression (amount) of the 110 kDa subunit of TFIIIC (121, 122), yet the actual ability of induced TFIIIC110 expression to modulate Pol III transcription could not be confirmed by a subsequent study (113). Analysis of the RNA-seq data in this study revealed: no significant variation of any TFIIIC subunit gene expression; a 2-fold increase for BRF1 expression; a significant increase in the expression of several Pol III subunit genes, in particular POLR3K/*RPC11* (4.8-fold)

and POLR3G/*RPC32a* (3.7-fold); a down-regulation (0.49-fold) of the Pol III subunit gene POLR3GL/*RPC32b.* These variations were calculated only considering the 24-hr mock and dl1500 samples.

Remarkably, POLR3G/*RPC32a* has been shown to be normally expressed at low levels in differentiated tissues, and to be increased at both mRNA and protein levels during IMR90 cell transformation (123).

Since a recent study (124) showed that POLR3G promoter contains a Myc-binding site, like all Pol III subunit genes except POLR3GL, we asked if e1a induces Myc expression. However looking at the data, it seems that MYC is instead 3.3-fold downregulated.

As mentioned above, it is worth noting that since we did not have any biological replicates, data reported for differential gene expression couldn't be subjected to any statistical test (see Supplementary Materials). We though tried to calculate differential genes expression using the mock and dl1500 samples 6 hr p.i. as a biological replicates of the 24 hr p.i. samples obtaining similar results.

4.4. *Alu* expression profiling in cancer cells

The ability to profile *Alu* expression at single-locus resolution opens novel interesting possibilities. As reminded above, a remarkable feature of *Alu* RNA profiles is that they most likely reflects the operation of epigenetic switches at the corresponding genomic loci. *Alu* expression profiles at single-locus resolution might lead to identify regions of particularly permissive chromatin whose resident genes might be deregulated in concomitance with *Alu* deregulation. More generally, *Alu* RNA profiles might represent a novel type of highly specific molecular signature for cancer and other diseases.

Of particular interest for this kind of approach are the RNA-seq data in The Cancer Genome Atlas (TCGA) (125), from which it will be possible to generate and comparatively analyze *Alu* expression profiles of thousands of human tumor samples.

TCGA is a comprehensive and coordinated effort to accelerate our understanding of the molecular basis of cancer through the application of genome analysis technologies, including large-scale genome sequencing.

Recently, TCGA RNA-seq datasets have been successfully exploited to investigate cancer-related alterations in pseudogene expression, alternative polyadenylation and promoter-associated small RNAs (126-128). With respect to these RNA profiling targets, *Alu* expression profiles, being subjected to strong epigenetic influence, have the potential to more directly and precisely reveal altered epigenomic states that might accompany malignancies. As an example of this possibility, demethylation (and thus likely derepression) at an *Alu* locus has been shown to activate an associated miRNA locus producing a miRNA acting as an oncogene endogenous silencer (129).

4.4.1. Preliminary results

As mentioned above the success of our bioinformatic pipeline to identify expression-positive *Alus* depends mostly on the RNA-seq data type which should contain sufficiently long sequencing reads to avoid excessive multimapping bias, preferably paired end and with strand information.

After a preliminary survey, we found that one of the TCGA RNA-seq datasets that most closely matches these technical specifications is the one relative to Stomach Adenocarcinoma (STAD) or gastric adenocarcinoma, whose molecular characterization has been extensively performed in a recent TCGA study (85).
Remarkably, in addition to providing data for a total of 416 STAD tumors, this dataset also contains matched tumor and non-tumor samples from 36 patients. We thus decided to focus on these samples, because they would have allowed us to test the feasibility of our approach by carrying out a focused comparative analysis of *Alu* expression profiles in matched cancer/normal tissues .

The 75-nt long unstranded paired-end RNA-seq reads of each sample have been aligned to the human genome (GRCh39/hg19) using TopHat and each bam file has been submitted to our in house developed *SINEsFind* Python script for genuine Pol III transcribed *Alu* RNA identification among the intergenic annotated *Alu* elements (see Supplementary Materials).
We found 5415 unique (i.e. counted once) expression-positive *Alus* among all the non-tumor samples of the 36 patients analyzed, while those in tumor samples were 14906 among which 1338 were in common with the non-tumor ones as reported in Table 8. Besides the almost 3-fold increase in the number of transcriptionally active *Alu* elements in the tumor with respect to non-tumor samples we saw a great variability among patients and even between the 2 conditions in the same patient. The full list of expression-positive *Alus*, along with their normalized expression values, are reported in Supplementary Table S12.

Table 8. Summary of number of expression-positive *Alus* in STAD TCGA dataset with their expression values (see Supplementary Materials for details)

	Alus		**Normalized expression coverage**	
Patient	**non-tumor**	**tumor**	**non-tumor**	**tumor**
1	327	382	1031011	897351
2	304	65	715066	418389
3	30	220	121518	424428
4	412	300	633854	474865
5	140	29	291064	203124
6	25	254	249520	605025
7	40	74	76468	173443
8	772	522	823001	1007461
9	633	737	1227017	699586
10	22	56	122732	154818
11	63	54	245919	92099
12	361	5988	933076	8035206
13	70	1569	380574	2580649
14	46	152	110332	410113
15	89	48	232409	171011
16	361	473	706744	1500042
17	91	82	269426	278196
18	107	3320	579876	5447431
19	46	79	92477	340972
20	25	69	88151	198933
21	91	38	195585	122633
22	257	1143	648996	1899439
23	42	147	104381	359408
24	63	92	154896	180461
25	114	107	112469	331475
26	48	493	84330	845843
27	1354	1322	3360956	2044449
28	40	54	89714	180336
29	56	35	87465	145327
30	327	437	1045365	965454
31	45	40	106417	122089
32	165	514	411097	766608
33	295	38	973487	279826
34	93	52	289884	381025
35	41	42	113541	124571
36	122	164	259770	469144
Total unique	**5415**	**14906**	**226248**	**444416**

We thus asked if this variability in the number of expression-positive *Alus* also reflected a variability of their expression levels. For each patient, since an expression-positive *Alu* could have been filtered out in a sample by our bioinformatic pipeline due to an expression coverage value under a pre-set background value (see Supplementary Materials), we decided to calculate expression levels in the tumor and non-tumor sample for all the *Alus* (counted once) found expression-positive in either conditions (see Supplementary Materials). As shown in Table 8, when the expression levels for all expression-positive *Alus* in the 36 patients are summed up, a roughly 2-fold increase in *Alu* expression levels is observed in the tumor with respect to non-tumor samples. We also asked whether this variability in the number of expressed *Alus* among patients and between conditions could have been arisen from the different sequencing depth of the samples but we concluded it was not the case.

In summary, these preliminary results support the notion that gastric adenocarcinoma is characterized not only by a general increase in the expression of *Alus* already transcriptionally active in non-tumor sample, but also by induction of transcription of new *Alu* loci. These results are in agreement with those in *dl*1500 AD5 infected IMR90 cells that show that e1a oncoprotein induce Pol III *Alu* transcription.

5. Discussion

The studies reported in this thesis provide the first comprehensive account of transcriptionally active *Alu* and MIR loci in human cells, reveal the existence of novel Pol III-transcribed genes originated from monomeric *Alu* elements and MIR fragments, and support the notion that SINE expression in human cells occurs rarely, from small, largely cell-specific sets of transcriptionally active SINEs regulated by both internal and external *cis*-acting control elements. They also strengthen the notion that *Alu* expression increases during cell stress such as viral infection and in cancer cells.

Historically, the tasks of detecting genuine Pol III-transcribed SINE RNAs and of attributing them to individual transcriptionally active SINE loci had to face two challenges: the extremely high copy number and sequence similarity of *Alu* and MIR elements within the human genome, and their frequent location within introns or untranslated regions of primary or mature Pol II transcripts. Previous studies of *Alu* expression exploited Northern hybridization, producing information on transcript size, as a useful tool in distinguishing genuine (~300-500 nt) *Alu* Pol III transcripts from *Alu* RNA incorporated into longer Pol II transcripts (even though probe cross-hybridization with the closely related, ~300 nt-long 7SL RNA might frequently represent a problem) (21, 130). Distinguishing between the products of individual *Alu* elements, or even of different *Alu* subfamilies, however, is unfeasible through Northern blot. *Alu* RNA detection approaches based on RT-PCR are even less effective in distinguishing genuine Pol III *Alu* transcripts from *Alu* RNA sequences included into Pol II-synthesized hnRNA or mRNA (21). To date, the only low-throughput approach that has permitted to identify genuine *Alu* Pol III transcripts, giving the possibility to trace the corresponding *Alu* loci, was based on a C-RACE technique (a modified version of a RACE which allow an unbiased isolation of 3' ends) disclosing sequence information on individual *Alu* RNAs (97). The recent development of unbiased genome-wide location analyses exploiting next-generation sequencing (NGS) technologies has allowed the identification, through ChIP-seq approaches, of several SINE loci that are bound *in vivo* by the Pol III transcription machinery, a reasonable indication of a transcriptionally active state (42, 70).

Within this context, the original contribution of our work proceeds from the simple remark that appropriate analysis of RNA-seq data, containing full sequence information even on rare transcripts, should allow to successfully face difficulties in both sequence and length determination of *Alu* and MIR transcripts. Indeed, by applying to ENCODE RNA-seq datasets an *ad hoc* devised computational search strategy, mainly relying on unique alignment and size-selection of RNA-seq signal mapping, we were able to unveil to an unprecedented detail the *Alu* and MIR transcriptomes of several human cell lines under different conditions. Our search algorithm appeared to work well especially for the identification of expressed intergenic/antisense SINE transcription units whose RNA products, in contrast to the ones located within Pol II genes in a sense orientation, tend to be less obscured by flanking unrelated RNA-seq signals. In strong support to the genuine nature of expression-positive intergenic/antisense *Alus* as independent Pol III transcription units is the observation that, in HeLa and K562 cell lines, a remarkable percentage of them (29% and 44%, respectively) was independently found to be bound by one or more components of the Pol III transcription machinery in independent ChIP-seq analyses. On the contrary these percentage fall down to ~20% and 27% for MIR transcription units which showed significant enrichment only for BDP1 and few other Pol III transcription components only in K562 cell line. Such a poor association of Pol III factors to expression-positive MIRs could eventually arise from a bias in the filtering option values used in the bioinformatics pipeline which have been set to a lower stringency as to include as much as possible putative Pol III transcribed MIR loci but including the possibility of more false positives. A modest overlap was also observed between our set of expression-positive intergenic *Alus* and the list of putative Pol III-transcribed *Alus* reported by a previous integrated analysis of ChIP seq studies of human Pol III machinery (70). A possible reason for this discrepancy could be the fact that, in contrast to that study, we also included in our analysis incomplete *Alu* elements that turned out to be contributing to expression-positive *Alu* set. Another possibility is that the compilation in (70) was based on partial lists of potentially transcribed *Alus* that had already been pre-selected by the authors of the different ChIP-seq studies according to very stringent criteria, which could have led to the exclusion of expression-positive *Alus*.

The most evident features of *Alu* and MIR expression profiles as revealed by our analysis are: i) the extremely low number of detectably expressed SINEs in each cell line, in the order of hundreds, corresponding to less than 0.1% of all annotated *Alus* and MIRs; ii) the existence, among intergenic/antisense expression-positive *Alus*, of an unexpectedly large set of elements expressed in more than one cell line,, suggesting that, in human cells, *Alu* transcript profiles result from the combined activities of very few transcription-prone *Alu* elements, that are thus reminiscent of the rare and elusive 'source' *Alu* elements possibly contributing to *Alu* expansion through retrotransposition (21); iii) even though different cell lines share a significant number of expression-positive *Alus*,, a marked cell-specificity of *Alu* and MIR transcriptomes is observed, thus suggesting that the *Alu* RNA expression profile in each cell line results from the expression of both commonly expressed and cell-specific *Alu* transcription units, while MIR RNA expression results mostly from cell specific expression units, possibly due to the above mentioned bias in the option values set in the bioinformatics pipeline; iv) *Alu* and MIR transcriptomes as revealed by ENCODE RNA-seq data analysis are composed of both full-length and incomplete SINE transcripts, some of which might be related to the previously described sc*Alu* transcripts corresponding to the left *Alu* monomer (with the caveat that *Alu* RNA fragment detection in our case might also result from non-physiological RNA degradation).

An interesting outcome of our analysis is the identification of novel monomeric *Alu* elements whose RNA-seq signal profiles suggest a transcription unit organization similar to the one firstly reported for the BC200 RNA gene (101): a promoter-containing *Alu* left monomer directing Pol III to synthesize a ncRNA containing the *Alu* sequence itself followed by an *Alu*-unrelated RNA moiety. The so-generated, *Alu*-derived ncRNAs have the potential to play novel regulatory roles deriving from the combination of an *Alu* left arm with unique RNA sequences. An *Alu* left monomer-derived gene that we find of particular interest, and that we have called Ya5-lm, is located in multiple copies on chromosome 19, with each copy located very close to one of the *snaR* gene copies belonging to either of two *snaR* clusters on chromosome 19 (102). Such a close spatial relationship between Ya5-lm and *snaR* genes (that also likely evolved from *Alu* left monomers) suggests that Ya5-lms have been included in the same segmental duplication through which *snaR* genes are thought to have spread. The *snaR* clusters on chromosome 19 might thus host a

chromatin environment favourable to Pol III transcription of different *Alu*-derived ncRNAs, possibly playing recently evolved functions in translation regulation (102, 131).

In a similar way we found that full-length ~280 nt long MIR transcripts may arise from incomplete MIR elements annotated onto the human genome and corresponding to either left or right fragment of canonical MIR sequence. In these cases, transcription appears either to initiate in an upstream MIR-unrelated region containing Pol III promoters or to continue in a downstream MIR-unrelated region, until the encounter of the Pol III terminator, producing chimeric transcripts similar to the Ya5-lm transcription unit. The regulation and possible function of these novel Pol III-transcribed genes awaits further characterization.

Perhaps not surprisingly, given the relatively frequent occurrence of intronic nested genes in metazoan genomes (132), our data also suggest that a number of gene-hosted (and particularly intron-hosted), sense-oriented *Alus* and MIRs are likely to represent autonomous transcription units that are recognized by the Pol III machinery and thus transcribed independently from Pol II transcription of the host gene. The possible interplay between Pol II and Pol III transcription of host and nested genes is an issue deserving further investigation, especially in light of recent evidence for the involvement of a Pol III-Pol II switch in the insulator activity of a mouse B1 SINE (133), and of the widespread association of Pol II factors with Pol III transcribed genes (104), including *Alus* and MIRs as clearly confirmed by our results (see Supplementary Table S3 and S7). Related to this issue is the observation that gene-hosted sense-oriented *Alus* and MIRs, revealed as expression-positive by our analysis, have a lesser tendency than intergenic/antisense ones to be associated with the Pol III machinery. This leads to speculate that the synthesis of gene-hosted (mostly intron-hosted) SINEs might occur either via the release of SINE RNAs from annotated Pol II-synthesized host transcripts [similarly to intron-derived microRNAs or snoRNAs (92, 93)], or through the still uncharacterized production and processing of unannotated *Alu/MIR*-containing noncoding Pol II transcripts possibly related to *Alu*-associated Pol II and TFs revealed by ChIP-seq analyses. This possibility also applies to intergenic/antisense *Alus* and MIRs found to be expression-positive but not Pol III-associated.

With respect to mechanistic understanding of *Alu* and MIR transcription and their control, our studies, by comparing *in vivo* expression levels with *in vitro*

transcription rates of a number of *Alu* and MIR loci, confirm and extend previous knowledge about two peculiar features of *Alu* and MIR transcription by Pol III: i) the stimulatory role of 5'-flanking sequences on *Alu*/MIR transcription; ii) the strong epigenetic control on *Alu* and MIR expression *in vivo*. Of the 9 *Alus* and 4 MIRs whose transcription properties were analyzed *in vitro* in the present study 6 *Alus* and 3 MIRs exhibited a ~2-fold or higher reduction of transcription upon deletion of the 5'-flanking region, while only one *Alu* was unaffected and 1 MIR showed moderately reduced expression.

That upstream sequences may influence transcription by Pol III of its target genes, even when they display internal promoters, is a well-documented possibility. For example tRNA genes, whose internal promoter organization closely resembles the one found in *Alus*, tend to display a certain degree of upstream sequence conservation in the genomes of different eukaryotic lineages and, correspondingly, their transcription appears to be influenced by upstream sequence context both *in vitro* and *in vivo* (134). In the case of *Alus*, the internal Pol III promoter has been suggested not to be sufficiently strong to warrant their efficient transcription independently from favourable upstream sequences (21). With this respect, *Alus* resemble their 7SL progenitor, whose sub-optimal internal promoter requires upstream sequence elements to direct efficient transcription (135). If the general consensus sequences for A- and B-boxes, mainly deduced from tRNA gene sequence analysis (TRGYnnAnnnG and GWTCRAnnC, respectively (27)) are compared with the highly conserved *Alu* A- and B-box sequences (TGGCTCACGCC and GWTCGAGAC (136)), a noticeable difference appears at the last position of the A box, which in *Alu* is C instead of G. Another difference is the distance between A and B boxes (50 and 35 bp in the case of *Alu* and of MIR and tRNA genes, respectively). Both of these peculiar features might contribute to the intrinsic weakness of *Alu* internal promoter, especially if one considers that A box acts as a fundamental core promoter element in Pol III transcription, frequently in synergy with upstream elements (27, 137). Interestingly, the A box (TGGCGCGTGCC) and B box (GTTCTGGGC) recognizable within the human 7SL genes differ from tRNA gene consensus even more than *Alu* internal control regions do, in line with the severe requirement of upstream control elements in 7SL gene transcription (135, 138).

Moreover even if the expression level of MIRs with mutationally inactivated B-box dramatically decreased, an appreciable levels of residual transcription was observed. This observation suggests the possibility that Pol III transcription of MIRs does not rely entirely on this element but could be supported to similar extent by other factors (e.g. 5'-flanking sequence) such that only mutating both together would be abolished.

The existence of a strong epigenetic control on SINE expression *in vivo* has previously been proposed and widely accepted to explain the discrepancy between the extremely high number of genomic *Alus* and MIR and the paucity of their overall expression level (reviewed in (49, 71)). In our study of ENCODE cell lines, *in vivo* epigenetic silencing can be easily deduced from the similar *in vitro* transcription rates of *Alu* and MIR elements which profoundly differ from each other for their expression properties in cell lines. DNA methylation is generally proposed as the main factor responsible for widespread SINE downregulation (49), which may also involve H3K9 methylation (139), even though more recent investigation on *Alu* histone modification patterns, based on ChIP-seq, revealed that, somehow unexpectedly, *Alus* tend to possess histone modifications (such as H3K4me1/2) generally associated with open chromatin and enhancers (52). Clearly, we are still missing important information on the mechanisms of general SINE silencing and local derepression and their relationship with DNA methylation and chromatin organization. An initial contribution to this issue is represented by our finding that the P300 histone acetyltransferase is enriched at expressed *Alu* loci, whose upstream regions also tend to be associated with JunD and C/EBP beta transcription factors in the investigated ENCODE cell lines. Similarly, expressed MIR loci where also found to be enriched, in 3 of the 7 ENCODE cell lines studied, in P300, along with TBP, MAZ, YY1 and PML transcription factors. In principle these Pol II TFs, that were found enriched in the 500 bp region upstream of expression-positive *Alus* and MIR, might be involved in the modulation of their Pol III transcription by 5'-flanking region, and they might also possibly favour hypothetical Pol II transcription at SINE loci which might contribute to *Alu* and MIR RNA biogenesis.

The identification of SINE expression profiles within introns of either coding or non-coding Pol II genes in both sense and antisense orientation confirms the ability of

the Pol III transcription machinery to access these loci. Moreover the observed SINE transcription antisense to the hosting gene, such as the well-documented MIR in the first intron of POLR3E, raises the possibility of an underlying mechanism for Pol II gene regulation by *Alus* and MIRs transcriptional units.

The established role of cellular stress signals in the activation of Pol III *Alus* transcription has been confirmed by our study showing that viral infection both induces transcription of new *Alu* loci and increases the expression of those already transcriptionally active in a time dependent manner. The enrichment in histones acetylation at expressed *Alu* loci in both mock and dl1500 infected cells support further the hypothesis of epigenetic control on *Alus* transcription. Moreover the known role of e1a oncoprotein in the activation of the TFIIIC complex by the induction of the TFIIIC-110 subunits is not supported by our data, which however show significative enrichment of this factor at expression-positive *Alu* loci.

Very little is known about the expression of Pol III-derived *Alu* RNAs in cancer cells and the majority of previous studies did not distinguish between them and *Alu*-containing transcripts arising from Pol II transcription. Even when such an effort was carried on trying to quantify the expression of genuine Pol III RNAs in hepatocellular carcinoma tissues (140) no information were available related to the source loci of transcription making it impossible to make correlation with other genome-wide data such as differential genes expression and other epigenetic deregulations. Indeed most of the evidences pointing to an increased expression of *Alu* RNAs are only theoretical and mostly based on the evidence of an increased DNA hypomethylation at these loci, while a more recent study showed that SINE transcription by RNA polymerase III is suppressed by histone methylation but not by DNA methylation (65).
Here we show how the observed increase in *Alu* expression following an *in vitro* pro-oncogenic stimulus (i.e. e1a overexpression) finds a corresponding result *in vivo* where, for the first time, we performed a genuine Pol III *Alu* expression profiling through TCGA RNA-seq data of gastric adenocarcinoma cells allowing us to identify the source loci of Pol III transcription and showing that *Alu* RNAs are overexpressed in cancer cells, opening the intriguing possibility to use their expression profiling as a novel epigenetic marker for cancer biology.

6. Conclusion

The ability to determine SINE expression profiles at single-locus resolution represents a key step towards a better understanding of *Alu* and MIR transcriptional control, a largely unexplored issue in spite of its high relevance for human genome stability. The possible cellular functions of SINE RNAs are just starting to be discerned (141); it is thus presently difficult to interpret SINE RNA profiles in terms of their significance in cell physiology. However, as suggested by the present works together with previous studies, SINE RNA profiles are likely determined by a tiny subset of loci particularly responsive to DNA methylation and chromatin status. SINE RNA profiling through RNA-seq might thus represent a novel, extremely subtle and sensitive way to monitor epigenome alterations accompanying physiological and pathological states. Our works open the possibility to easily profile the human transcriptome in any human cell line or tissue, under any condition compatible with RNAseq. We anticipate that our pipeline will be widely exploited to extract unprecedented information on SINE expression profiles from the plethora of available human RNA-seq datasets. Of particular interest with this respect will be SINE RNA profiling in relation to development, malignant transformation, cellular alteration in various diseases, and inter-individual differences in gene expression.

Part of this work has already been published in international journals (142, 143).

References

1. McClintock B. Controlling elements and the gene. Cold Spring Harbor symposia on quantitative biology. 1956;21:197-216.
2. Britten RJ, Davidson EH. Gene regulation for higher cells: a theory. Science. 1969;165(3891):349-57.
3. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, et al. Initial sequencing and analysis of the human genome. Nature. 2001;409(6822):860-921.
4. Deininger PL, Moran JV, Batzer MA, Kazazian HH, Jr. Mobile elements and mammalian genome evolution. Current opinion in genetics & development. 2003;13(6):651-8.
5. Goodier JL, Kazazian HH, Jr. Retrotransposons revisited: the restraint and rehabilitation of parasites. Cell. 2008;135(1):23-35.
6. Kolosha VO, Martin SL. In vitro properties of the first ORF protein from mouse LINE-1 support its role in ribonucleoprotein particle formation during retrotransposition. Proceedings of the National Academy of Sciences of the United States of America. 1997;94(19):10155-60.
7. Martin SL, Cruceanu M, Branciforte D, Wai-Lun Li P, Kwok SC, Hodges RS, et al. LINE-1 retrotransposition requires the nucleic acid chaperone activity of the ORF1 protein. Journal of molecular biology. 2005;348(3):549-61.
8. Feng Q, Moran JV, Kazazian HH, Jr., Boeke JD. Human L1 retrotransposon encodes a conserved endonuclease required for retrotransposition. Cell. 1996;87(5):905-16.
9. Mathias SL, Scott AF, Kazazian HH, Jr., Boeke JD, Gabriel A. Reverse transcriptase encoded by a human transposable element. Science. 1991;254(5039):1808-10.
10. Wei W, Gilbert N, Ooi SL, Lawler JF, Ostertag EM, Kazazian HH, et al. Human L1 retrotransposition: cis preference versus trans complementation. Molecular and cellular biology. 2001;21(4):1429-39.
11. Prak ET, Kazazian HH, Jr. Mobile elements and the human genome. Nature reviews Genetics. 2000;1(2):134-44.
12. Beck CR, Garcia-Perez JL, Badge RM, Moran JV. LINE-1 elements in structural variation and disease. Annual review of genomics and human genetics. 2011;12:187-215.
13. Brouha B, Schustak J, Badge RM, Lutz-Prigge S, Farley AH, Moran JV, et al. Hot L1s account for the bulk of retrotransposition in the human population. Proceedings of the National Academy of Sciences of the United States of America. 2003;100(9):5280-5.
14. Cordaux R, Batzer MA. The impact of retrotransposons on human genome evolution. Nature reviews Genetics. 2009;10(10):691-703.
15. Eickbush TH. Transposing without ends: the non-LTR retrotransposable elements. The New biologist. 1992;4(5):430-40.
16. Kramerov DA, Vassetzky NS. SINEs. Wiley interdisciplinary reviews RNA. 2011;2(6):772-86.
17. Richardson SR, Doucet AJ, Kopera HC, Moldovan JB, Garcia-Perez JL, Moran JV. The Influence of LINE-1 and SINE Retrotransposons on Mammalian Genomes. Microbiology spectrum. 2015;3(2):MDNA3-0061-2014.
18. Oler AJ, Alla RK, Roberts DN, Wong A, Hollenhorst PC, Chandler KJ, et al. Human RNA polymerase III transcriptomes and relationships to Pol II promoter chromatin and enhancer-binding factors. Nature structural & molecular biology. 2010;17(5):620-8.

19. Varshney D, Vavrova-Anderson J, Oler AJ, Cairns BR, White RJ. Selective repression of SINE transcription by RNA polymerase III. Mobile Genetic Elements. 2015;5(6):86-91.
20. Shen MR, Batzer MA, Deininger PL. Evolution of the master Alu gene(s). Journal of molecular evolution. 1991;33(4):311-20.
21. Deininger P. Alu elements: know the SINEs. Genome biology. 2011;12(12):236.
22. Dewannieux M, Heidmann T. Role of poly(A) tail length in Alu retrotransposition. Genomics. 2005;86(3):378-81.
23. Roy-Engel AM, Salem AH, Oyeniran OO, Deininger L, Hedges DJ, Kilroy GE, et al. Active Alu element "A-tails": size does matter. Genome research. 2002;12(9):1333-44.
24. Dieci G, Conti A, Pagano A, Carnevali D. Identification of RNA polymerase III-transcribed genes in eukaryotic genomes. Biochimica et biophysica acta. 2013;1829(3-4):296-305.
25. Elder JT, Pan J, Duncan CH, Weissman SM. Transcriptional analysis of interspersed repetitive polymerase III transcription units in human DNA. Nucleic acids research. 1981;9(5):1171-89.
26. Fuhrman SA, Deininger PL, LaPorte P, Friedmann T, Geiduschek EP. Analysis of transcription of the human Alu family ubiquitous repeating element by eukaryotic RNA polymerase III. Nucleic acids research. 1981;9(23):6439-56.
27. Orioli A, Pascali C, Pagano A, Teichmann M, Dieci G. RNA polymerase III transcription control elements: themes and variations. Gene. 2012;493(2):185-94.
28. Chesnokov I, Schmid CW. Flanking sequences of an Alu source stimulate transcription in vitro by interacting with sequence-specific transcription factors. Journal of molecular evolution. 1996;42(1):30-6.
29. Roy AM, West NC, Rao A, Adhikari P, Aleman C, Barnes AP, et al. Upstream flanking sequences and transcription of SINEs. Journal of molecular biology. 2000;302(1):17-25.
30. Paolella G, Lucero MA, Murphy MH, Baralle FE. The Alu family repeat promoter has a tRNA-like bipartite structure. The EMBO journal. 1983;2(5):691-6.
31. Orioli A, Pascali C, Quartararo J, Diebel KW, Praz V, Romascano D, et al. Widespread occurrence of non-canonical transcription termination by human RNA polymerase III. Nucleic acids research. 2011;39(13):5499-512.
32. West N, Roy-Engel AM, Imataka H, Sonenberg N, Deininger P. Shared protein components of SINE RNPs. Journal of molecular biology. 2002;321(3):423-32.
33. Hsu K, Chang DY, Maraia RJ. Human signal recognition particle (SRP) Alu-associated protein also binds Alu interspersed repeat sequence RNAs. Characterization of human SRP9. The Journal of biological chemistry. 1995;270(17):10179-86.
34. Ade C, Roy-Engel AM, Deininger PL. Alu elements: an intrinsic source of human genome instability. Current opinion in virology. 2013;3(6):639-45.
35. Wallace N, Wagstaff BJ, Deininger PL, Roy-Engel AM. LINE-1 ORF1 protein enhances Alu SINE retrotransposition. Gene. 2008;419(1-2):1-6.
36. Jurka J, Zietkiewicz E, Labuda D. Ubiquitous mammalian-wide interspersed repeats (MIRs) are molecular fossils from the mesozoic era. Nucleic acids research. 1995;23(1):170-5.
37. Smit AF, Riggs AD. MIRs are classic, tRNA-derived SINEs that amplified before the mammalian radiation. Nucleic acids research. 1995;23(1):98-102.

38. Terai Y, Takahashi K, Okada N. SINE cousins: the 3'-end tails of the two oldest and distantly related families of SINEs are descended from the 3' ends of LINEs with the same genealogical origin. Molecular biology and evolution. 1998;15(11):1460-71.
39. Chalei MB, Korotkov EV. Evolution of MIR Elements Located in the Coding Regions of Human Genome. Molekuliarnaia biologiia. 2001;35(6):1023-31.
40. Sironi M, Menozzi G, Comi GP, Cereda M, Cagliani R, Bresolin N, et al. Gene function and expression level influence the insertion/fixation dynamics of distinct transposon families in mammalian introns. Genome biology. 2006;7(12):R120.
41. Jjingo D, Conley AB, Wang J, Marino-Ramirez L, Lunyak VV, Jordan IK. Mammalian-wide interspersed repeat (MIR)-derived enhancers and the regulation of human gene expression. Mobile DNA. 2014;5:14.
42. Canella D, Praz V, Reina JH, Cousin P, Hernandez N. Defining the RNA polymerase III transcriptome: Genome-wide localization of the RNA polymerase III transcription machinery in human cells. Genome research. 2010;20(6):710-21.
43. Carriere L, Graziani S, Alibert O, Ghavi-Helm Y, Boussouar F, Humbertclaude H, et al. Genomic binding of Pol III transcription machinery and relationship with TFIIS transcription factor distribution in mouse embryonic stem cells. Nucleic acids research. 2012;40(1):270-83.
44. Canella D, Bernasconi D, Gilardi F, LeMartelot G, Migliavacca E, Praz V, et al. A multiplicity of factors contributes to selective RNA polymerase III occupancy of a subset of RNA polymerase III genes in mouse liver. Genome research. 2012;22(4):666-80.
45. Luan DD, Korman MH, Jakubczak JL, Eickbush TH. Reverse transcription of R2Bm RNA is primed by a nick at the chromosomal target site: a mechanism for non-LTR retrotransposition. Cell. 1993;72(4):595-605.
46. Kulpa DA, Moran JV. Cis-preferential LINE-1 reverse transcriptase activity in ribonucleoprotein particles. Nature structural & molecular biology. 2006;13(7):655-60.
47. Levin HL, Moran JV. Dynamic interactions between transposable elements and their hosts. Nature reviews Genetics. 2011;12(9):615-27.
48. Kroutter EN, Belancio VP, Wagstaff BJ, Roy-Engel AM. The RNA polymerase dictates ORF1 requirement and timing of LINE and SINE retrotransposition. PLoS genetics. 2009;5(4):e1000458.
49. Roy-Engel AM. LINEs, SINEs and other retroelements: do birds of a feather flock together? Frontiers in bioscience. 2012;17:1345-61.
50. Goodier JL, Zhang L, Vetter MR, Kazazian HH, Jr. LINE-1 ORF1 protein localizes in stress granules with other RNA-binding proteins, including components of RNA interference RNA-induced silencing complex. Molecular and cellular biology. 2007;27(18):6469-83.
51. Rebollo R, Romanish MT, Mager DL. Transposable elements: an abundant and natural source of regulatory sequences for host genes. Annual review of genetics. 2012;46:21-42.
52. Su M, Han D, Boyd-Kirkup J, Yu X, Han JD. Evolution of Alu elements toward enhancers. Cell reports. 2014;7(2):376-85.
53. Wang J, Vicente-Garcia C, Seruggia D, Molto E, Fernandez-Minan A, Neto A, et al. MIR retrotransposon sequences provide insulators to the human genome. Proceedings of the National Academy of Sciences of the United States of America. 2015;112(32):E4428-37.

54. Hancks DC, Kazazian HH, Jr. Active human retrotransposons: variation and disease. Current opinion in genetics & development. 2012;22(3):191-203.
55. Kazazian HH, Jr., Wong C, Youssoufian H, Scott AF, Phillips DG, Antonarakis SE. Haemophilia A resulting from de novo insertion of L1 sequences represents a novel mechanism for mutation in man. Nature. 1988;332(6160):164-6.
56. Holdt LM, Hoffmann S, Sass K, Langenberger D, Scholz M, Krohn K, et al. Alu elements in ANRIL non-coding RNA at chromosome 9p21 modulate atherogenic cell functions through trans-regulation of gene networks. PLoS genetics. 2013;9(7):e1003588.
57. Gong C, Maquat LE. lncRNAs transactivate STAU1-mediated mRNA decay by duplexing with 3' UTRs via Alu elements. Nature. 2011;470(7333):284-8.
58. Chen LL, DeCerbo JN, Carmichael GG. Alu element-mediated gene silencing. The EMBO journal. 2008;27(12):1694-705.
59. Borchert GM, Lanier W, Davidson BL. RNA polymerase III transcribes human microRNAs. Nature structural & molecular biology. 2006;13(12):1097-101.
60. Heras SR, Macias S, Plass M, Fernandez N, Cano D, Eyras E, et al. The Microprocessor controls the activity of mammalian retrotransposons. Nature structural & molecular biology. 2013;20(10):1173-81.
61. Iwasaki YW, Siomi MC, Siomi H. PIWI-Interacting RNA: Its Biogenesis and Functions. Annual review of biochemistry. 2015;84:405-33.
62. Hu Q, Tanasa B, Trabucchi M, Li W, Zhang J, Ohgi KA, et al. DICER- and AGO3-dependent generation of retinoic acid-induced DR2 Alu RNAs regulates human stem cell proliferation. Nature structural & molecular biology. 2012;19(11):1168-75.
63. Friedli M, Trono D. The Developmental Control of Transposable Elements and the Evolution of Higher Species. Annual review of cell and developmental biology. 2015;31:429-51.
64. Chen L, Dahlstrom JE, Lee SH, Rangasamy D. Naturally occurring endo-siRNA silences LINE-1 retrotransposons in human cells through DNA methylation. Epigenetics. 2012;7(7):758-71.
65. Varshney D, Vavrova-Anderson J, Oler AJ, Cowling VH, Cairns BR, White RJ. SINE transcription by RNA polymerase III is suppressed by histone methylation but not by DNA methylation. Nature communications. 2015;6:6569.
66. Soifer HS, Zaragoza A, Peyvan M, Behlke MA, Rossi JJ. A potential role for RNA interference in controlling the activity of the human LINE-1 retrotransposon. Nucleic acids research. 2005;33(3):846-56.
67. Yang N, Kazazian HH, Jr. L1 retrotransposition is suppressed by endogenously encoded small interfering RNAs in human cultured cells. Nature structural & molecular biology. 2006;13(9):763-71.
68. Bogerd HP, Wiegand HL, Hulme AE, Garcia-Perez JL, O'Shea KS, Moran JV, et al. Cellular inhibitors of long interspersed element 1 and Alu retrotransposition. Proceedings of the National Academy of Sciences of the United States of America. 2006;103(23):8780-5.
69. Dieci G, Bosio MC, Fermi B, Ferrari R. Transcription reinitiation by RNA polymerase III. Biochimica et biophysica acta. 2013;1829(3-4):331-41.
70. Oler AJ, Traina-Dorge S, Derbes RS, Canella D, Cairns BR, Roy-Engel AM. Alu expression in human cell lines and their retrotranspositional potential. Mobile DNA. 2012;3(1):11.
71. Ichiyanagi K. Epigenetic regulation of transcription and possible functions of mammalian short interspersed elements, SINEs. Genes & genetic systems. 2013;88(1):19-29.

72. Panning B, Smiley JR. Activation of RNA polymerase III transcription of human Alu repetitive elements by adenovirus type 5: requirement for the E1b 58-kilodalton protein and the products of E4 open reading frames 3 and 6. Molecular and cellular biology. 1993;13(6):3231-44.
73. Panning B, Smiley JR. Activation of expression of multiple subfamilies of human Alu elements by adenovirus type 5 and herpes simplex virus type 1. Journal of molecular biology. 1995;248(3):513-24.
74. Liu WM, Chu WM, Choudary PV, Schmid CW. Cell stress and translational inhibitors transiently increase the abundance of mammalian SINE transcripts. Nucleic acids research. 1995;23(10):1758-65.
75. Yakovchuk P, Goodrich JA, Kugel JF. B2 RNA and Alu RNA repress transcription by disrupting contacts between RNA polymerase II and promoter DNA within assembled complexes. Proceedings of the National Academy of Sciences of the United States of America. 2009;106(14):5569-74.
76. Li TH, Kim C, Rubin CM, Schmid CW. K562 cells implicate increased chromatin accessibility in Alu transcriptional activation. Nucleic acids research. 2000;28(16):3031-9.
77. Laperriere D, Wang TT, White JH, Mader S. Widespread Alu repeat-driven expansion of consensus DR2 retinoic acid response elements during primate evolution. BMC genomics. 2007;8:23.
78. Agarwal P, Enroth S, Teichmann M, Wiklund HJ, Smit A, Westermark B, et al. Growth signals employ CGGBP1 to suppress transcription of Alu-SINEs. Cell cycle. 2014:0.
79. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nature methods. 2008;5(7):621-8.
80. Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. Nature reviews Genetics. 2009;10(1):57-63.
81. Treangen TJ, Salzberg SL. Repetitive DNA and next-generation sequencing: computational challenges and solutions. Nature reviews Genetics. 2012;13(1):36-46.
82. Montell C, Fisher EF, Caruthers MH, Berk AJ. Resolving the functions of overlapping viral genes by site-specific mutagenesis at a mRNA splice site. Nature. 1982;295(5848):380-4.
83. Ferrari R, Gou D, Jawdekar G, Johnson SA, Nava M, Su T, et al. Adenovirus small E1A employs the lysine acetylases p300/CBP and tumor suppressor Rb to repress select host genes and promote productive virus infection. Cell host & microbe. 2014;16(5):663-76.
84. Ferrari R, Su T, Li B, Bonora G, Oberai A, Chan Y, et al. Reorganization of the host epigenome by a viral oncogene. Genome research. 2012;22(7):1212-21.
85. Cancer Genome Atlas Research N. Comprehensive molecular characterization of gastric adenocarcinoma. Nature. 2014;513(7517):202-9.
86. Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome biology. 2013;14(4):R36.
87. Anders S, Pyl TP, Huber W. HTSeq — A Python framework to work with high-throughput sequencing data. . biorXiv preprint. 2014;doi: 10.1101/002824.
88. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013;29(1):15-21.

89. Preti M, Ribeyre C, Pascali C, Bosio MC, Cortelazzi B, Rougemont J, et al. The telomere-binding protein Tbf1 demarcates snoRNA gene promoters in Saccharomyces cerevisiae. Molecular cell. 2010;38(4):614-20.
90. Dignam JD, Lebovitz RM, Roeder RG. Accurate transcription initiation by RNA polymerase II in a soluble extract from isolated mammalian nuclei. Nucleic acids research. 1983;11(5):1475-89.
91. Djebali S, Davis CA, Merkel A, Dobin A, Lassmann T, Mortazavi A, et al. Landscape of transcription in human cells. Nature. 2012;489(7414):101-8.
92. Dieci G, Preti M, Montanini B. Eukaryotic snoRNAs: a paradigm for gene expression flexibility. Genomics. 2009;94(2):83-8.
93. Hesselberth JR. Lives that introns lead after splicing. Wiley interdisciplinary reviews RNA. 2013;4(6):677-91.
94. Liu WM, Maraia RJ, Rubin CM, Schmid CW. Alu transcripts: cytoplasmic localisation and regulation by DNA methylation. Nucleic acids research. 1994;22(6):1087-95.
95. Comeaux MS, Roy-Engel AM, Hedges DJ, Deininger PL. Diverse cis factors controlling Alu retrotransposition: what causes Alu elements to die? Genome research. 2009;19(4):545-55.
96. Maraia RJ, Driscoll CT, Bilyeu T, Hsu K, Darlington GJ. Multiple dispersed loci produce small cytoplasmic Alu RNA. Molecular and cellular biology. 1993;13(7):4233-41.
97. Shaikh TH, Roy AM, Kim J, Batzer MA, Deininger PL. cDNAs derived from primary and small cytoplasmic Alu (scAlu) transcripts. Journal of molecular biology. 1997;271(2):222-34.
98. Kojima KK. Alu monomer revisited: recent generation of Alu monomers. Molecular biology and evolution. 2011;28(1):13-5.
99. Sassa T. The role of human-specific gene duplications during brain development and evolution. Journal of neurogenetics. 2013;27(3):86-96.
100. Parrott AM, Mathews MB. snaR genes: recent descendants of Alu involved in the evolution of chorionic gonadotropins. Cold Spring Harbor symposia on quantitative biology. 2009;74:363-73.
101. Martignetti JA, Brosius J. BC200 RNA: a neural RNA polymerase III product encoded by a monomeric Alu element. Proceedings of the National Academy of Sciences of the United States of America. 1993;90(24):11563-7.
102. Parrott AM, Tsai M, Batchu P, Ryan K, Ozer HL, Tian B, et al. The evolution and expression of the snaR family of small non-coding RNAs. Nucleic acids research. 2011;39(4):1485-500.
103. Moqtaderi Z, Wang J, Raha D, White RJ, Snyder M, Weng Z, et al. Genomic binding profiles of functionally distinct RNA polymerase III transcription complexes in human cells. Nature structural & molecular biology. 2010;17(5):635-40.
104. White RJ. Transcription by RNA polymerase III: more complex than we thought. Nature reviews Genetics. 2011;12(7):459-63.
105. Donze D. Extra-transcriptional functions of RNA Polymerase III complexes: TFIIIC as a potential global chromatin bookmark. Gene. 2012;493(2):169-75.
106. Pascali C, Teichmann M. RNA polymerase III transcription - regulated by chromatin structure and regulator of nuclear chromatin organization. Sub-cellular biochemistry. 2013;61:261-87.

107. Raha D, Wang Z, Moqtaderi Z, Wu L, Zhong G, Gerstein M, et al. Close association of RNA polymerase II and many transcription factors with Pol III genes. Proceedings of the National Academy of Sciences of the United States of America. 2010;107(8):3639-44.
108. Percudani R, Pavesi A, Ottonello S. Transfer RNA gene redundancy and translational selection in Saccharomyces cerevisiae. Journal of molecular biology. 1997;268(2):322-30.
109. Athanikar JN, Badge RM, Moran JV. A YY1-binding site is required for accurate human LINE-1 transcription initiation. Nucleic acids research. 2004;32(13):3846-55.
110. Ernst J, Kellis M. ChromHMM: automating chromatin-state discovery and characterization. Nature methods. 2012;9(3):215-6.
111. Ferrari R, Pellegrini M, Horwitz GA, Xie W, Berk AJ, Kurdistani SK. Epigenetic reprogramming by adenovirus e1a. Science. 2008;321(5892):1086-8.
112. Hoeffler WK, Kovelman R, Roeder RG. Activation of transcription factor IIIC by the adenovirus E1A protein. Cell. 1988;53(6):907-20.
113. Innes F, Ramsbottom B, White RJ. A test of the model that RNA polymerase III transcription is regulated by selective induction of the 110 kDa subunit of TFIIIC. Nucleic acids research. 2006;34(11):3399-407.
114. Kirkland JG, Raab JR, Kamakaka RT. TFIIIC bound DNA elements in nuclear organization and insulation. Biochimica et biophysica acta. 2013;1829(3-4):418-24.
115. Van Bortle K, Corces VG. tDNA insulators and the emerging role of TFIIIC in genome organization. Transcription. 2012;3(6):277-84.
116. Horwitz GA, Zhang K, McBrian MA, Grunstein M, Kurdistani SK, Berk AJ. Adenovirus small e1a alters global patterns of histone modification. Science. 2008;321(5892):1084-5.
117. Mertens C, Roeder RG. Different functional modes of p300 in activation of RNA polymerase III transcription from chromatin templates. Molecular and cellular biology. 2008;28(18):5764-76.
118. Deininger PL, Batzer MA. Alu repeats and human disease. Molecular genetics and metabolism. 1999;67(3):183-93.
119. Hoeffler WK, Roeder RG. Enhancement of RNA polymerase III transcription by the E1A gene product of adenovirus. Cell. 1985;41(3):955-63.
120. Gaynor RB, Feldman LT, Berk AJ. Transcription of class III genes activated by viral immediate early proteins. Science. 1985;230(4724):447-50.
121. Yoshinaga S, Dean N, Han M, Berk AJ. Adenovirus stimulation of transcription by RNA polymerase III: evidence for an E1A-dependent increase in transcription factor IIIC concentration. The EMBO journal. 1986;5(2):343-54.
122. Sinn E, Wang Z, Kovelman R, Roeder RG. Cloning and characterization of a TFIIIC2 subunit (TFIIIC beta) whose presence correlates with activation of RNA polymerase III-mediated transcription by adenovirus E1A expression and serum factors. Genes & development. 1995;9(6):675-85.
123. Haurie V, Durrieu-Gaillard S, Dumay-Odelot H, Da Silva D, Rey C, Prochazkova M, et al. Two isoforms of human RNA polymerase III with specific functions in cell growth and transformation. Proceedings of the National Academy of Sciences of the United States of America. 2010;107(9):4176-81.

124. Renaud M, Praz V, Vieu E, Florens L, Washburn MP, l'Hote P, et al. Gene duplication and neofunctionalization: POLR3G and POLR3GL. Genome Res. 2014;24(1):37-51.
125. Cancer Genome Atlas Research N, Weinstein JN, Collisson EA, Mills GB, Shaw KR, Ozenberger BA, et al. The Cancer Genome Atlas Pan-Cancer analysis project. Nature genetics. 2013;45(10):1113-20.
126. Han L, Yuan Y, Zheng S, Yang Y, Li J, Edgerton ME, et al. The Pan-Cancer analysis of pseudogene expression reveals biologically and clinically relevant tumour subtypes. Nature communications. 2014;5:3963.
127. Xia Z, Donehower LA, Cooper TA, Neilson JR, Wheeler DA, Wagner EJ, et al. Dynamic analyses of alternative polyadenylation from RNA-seq reveal a 3'-UTR landscape across seven tumour types. Nature communications. 2014;5:5274.
128. Zovoilis A, Mungall AJ, Moore R, Varhol R, Chu A, Wong T, et al. The expression level of small non-coding RNAs derived from the first exon of protein-coding genes is predictive of cancer status. EMBO reports. 2014;15(4):402-10.
129. Saito Y, Suzuki H, Tsugawa H, Nakagawa I, Matsuzaki J, Kanai Y, et al. Chromatin remodeling at Alu repeats by epigenetic treatment activates silenced microRNA-512-5p with downregulation of Mcl-1 in human gastric cancer cells. Oncogene. 2009;28(30):2738-44.
130. Chang DY, Maraia RJ. A cellular protein binds B1 and Alu small cytoplasmic RNAs in vitro. The Journal of biological chemistry. 1993;268(9):6423-8.
131. Eom T, Muslimov IA, Tsokas P, Berardi V, Zhong J, Sacktor TC, et al. Neuronal BC RNAs cooperate with eIF4B to mediate activity-dependent translational control. The Journal of cell biology. 2014.
132. Kumar A. An overview of nested genes in eukaryotic genomes. Eukaryotic cell. 2009;8(9):1321-9.
133. Roman AC, Gonzalez-Rico FJ, Molto E, Hernando H, Neto A, Vicente-Garcia C, et al. Dioxin receptor and SLUG transcription factors regulate the insulator activity of B1 SINE retrotransposons via an RNA polymerase switch. Genome research. 2011;21(3):422-32.
134. Giuliodori S, Percudani R, Braglia P, Ferrari R, Guffanti E, Ottonello S, et al. A composite upstream sequence motif potentiates tRNA gene transcription in yeast. Journal of molecular biology. 2003;333(1):1-20.
135. Ullu E, Weiner AM. Upstream sequences modulate the internal promoter of the human 7SL RNA gene. Nature. 1985;318(6044):371-4.
136. Batzer MA, Deininger PL. Alu repeats and human genomic diversity. Nature reviews Genetics. 2002;3(5):370-9.
137. Guffanti E, Ferrari R, Preti M, Forloni M, Harismendy O, Lefebvre O, et al. A minimal promoter for TFIIIC-dependent in vitro transcription of snoRNA and tRNA genes by RNA polymerase III. The Journal of biological chemistry. 2006;281(33):23945-57.
138. Englert M, Felis M, Junker V, Beier H. Novel upstream and intragenic control elements for the RNA polymerase III-dependent transcription of human 7SL RNA genes. Biochimie. 2004;86(12):867-74.
139. Kondo Y, Issa JP. Enrichment for histone H3 lysine 9 methylation at Alu repeats in human cells. The Journal of biological chemistry. 2003;278(30):27658-62.

140. Tang RB, Wang HY, Lu HY, Xiong J, Li HH, Qiu XH, et al. Increased level of polymerase III transcribed Alu RNA in hepatocellular carcinoma tissue. Molecular carcinogenesis. 2005;42(2):93-6.
141. Berger A, Strub K. Multiple Roles of Alu-Related Noncoding RNAs. Progress in molecular and subcellular biology. 2011;51:119-46.
142. Conti A, Carnevali D, Bollati V, Fustinoni S, Pellegrini M, Dieci G. Identification of RNA polymerase III-transcribed Alu loci by computational screening of RNA-Seq data. Nucleic acids research. 2015;43(2):817-35.
143. Carnevali D, Dieci G. Alu expression profiles as a novel RNA signature in biology and disease. RNA & Disease. 2015.

Supplementary Materials

Supplementary Methods

Alu ENCODE: datasets

For *Alu* RNA identification, we used the Cold Spring Harbor Lab (CSHL) long RNA-seq data within ENCODE (whole-cell PolyA+ and PolyA- RNAs, two replicates for each sample) relative to the following cell lines: Gm12878, H1-hESC, HeLa-S3, HepG2, HUVEC, K562, NHEK, for a total of 28 datasets (http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeCshlLongRnaSeq).

The *Alu* dataset was downloaded from UCSC table browser Repeatmask track. It thus included, in addition to full-length, also 5' and/or 3'-truncated *Alu* elements. Since the GTF file reports non-unique *Alu* ids, we modified it to also report the chromosome name in order to obtain unique ids.

Alu ENCODE: Bioinfoinformatic pipeline for individually expressed *Alu* identification

Reads from each dataset were aligned to the reference genome (hg19) using TopHat aligner version 2.0.11 (1):

```
$ tophat -p 2 --b2-very-sensitive -o output_dir –library-type fr-firststrand genome pe1.fastq pe2.fastq
```

For the more permissive 'best match' alignment strategy (see Results section; (2)) we supplied to TopHat the "–g 1" option. We then used the htseq-count script of the HTSeq Python package (3) to count, for each dataset, the number of paired-end reads ("units of evidence" of sequenced cDNA fragment, see http://www-huber.embl.de/users/anders/HTSeq/doc/count.html) that mapped uniquely to each annotated *Alu*:

```
$ htseq-count -i transcript_id -m intersection-nonempty -s reverse sam_file Alu.gtf > counts
```

Each count table produced by htseq-count was then filtered in order to retain only *Alu* ids with more than 10 reads ("units of evidence") mapped. This threshold was chosen on the basis of the following reasoning. We considered the purely hypothetical case in which all the uniquely mapped reads on hg19 (ranging from

83731494 to 181434481 in the different cell lines) are uniformly distributed onto both strands of the human genome. We then calculated what would be, in this case, the number of reads mapping to a 300 bp region (the *Alu* length). By dividing, for each dataset, the number of uniquely mapped reads by 6 billion (the stranded human genome size) and multiplying by 300, we obtained an average expression coverage ranging from 4.2 and 9 reads for a 300 nt region.

The coordinates of retained *Alu*s, in a BED file format, was supplied to sitepro script of the Cis-regulatory Element Annotation System (CEAS suite; http://liulab.dfci.harvard.edu/CEAS/) along with the corresponding RNA-seq stranded signal profiles (in wiggle format) computed as described below.

First the BAM file containing uniquely mapped reads (NH:I:1) was split into two parts, based on the origin of the transcript from which the reads come from (using the TopHat XS tag) using samtools (4):

```
$ samtools view -H uniquely_mapped_reads.bam > plus_uniquely_mapped_reads.sam
$ samtools view -h uniquely_mapped_reads.bam|grep 'XS:A:+'>>
plus_uniquely_mapped_reads.sam
$ samtools view -Sb plus_uniquely_mapped_reads.sam >
plus_uniquely_mapped_reads.bam
$ samtools view -H uniquely_mapped_reads.bam >
minus_uniquely_mapped_reads.sam
$ samtools view -h uniquely_mapped_reads.bam|grep 'XS:A:-'>>
minus_uniquely_mapped_reads.sam
$ samtools view -Sb minus_uniquely_mapped_reads.sam >
minus_uniquely_mapped_reads.bam
```

The coverage was then computed for each of these stranded BAM files, without strand information, using genomeCoverageBed from bedtools (http://bedtools.readthedocs.org/en/latest/index.html):

```
$ genomeCoverageBed -split -bg -ibam plus_uniquely_mapped_reads.bam
-g hg19 > plus_uniquely_mapped_reads.bedgraph

$ genomeCoverageBed -split -bg -ibam minus_uniquely_mapped_reads.bam -g hg19
> minus_uniquely_mapped_reads.bedgraph
```

Finally the bedgraph files of each dataset were converted to WIG format to be used with SitePro along with *Alu* dataset filtered by strand:

```
$ sitepro --span=500 --dump --name=plus -b plus_Alu.bed -w plus_uniquely_mapped_reads.wig -l plus
$ sitepro --span=500 --dump --name=minus -b minus_Alu.bed -w minus_uniquely_mapped_reads.wig -l minus
```

We used sitepro (developed mainly for ChIP data) because it allowed us to calculate the signal profile in a range of +/- 500 nt from the center of the *Alu*s body with a resolution of 50 nt. Using the dump files we divided the whole region of 1050 nt calculated by sitepro (+/- 500 from the center of the *Alu* body, plus the 50-nt bin in the center) in three regions of 350 nt each, to which we will refer as 5'segment, *Alu* body and 3'segment, and calculated for each of them the cumulative signal of the corresponding seven 50-nt bins. For each RNA-seq dataset we thus obtained two tables, one for each strand, reporting, for each of the retained *Alu*s, three values corresponding to signals of the three regions mentioned above.

Since *Alu*s frequently lie close to, or within, Pol II genes, or they may be passengers of longer unknown transcripts, we devised a filter to be applied to these tables in order to mostly identify independently accumulated *Alu* transcripts (most likely synthesized by Pol III).

Since the *Alu* transcription start site (TSS) is located ~12 nt upstream the A-box, to avoid background noise due to upstream flanking RNA, we imposed for the 5'segment a very low value, i.e. less than 1/7 of the *Alu* body value (to include the possibility that in some cases transcription might start slightly upstream of the predicted TSS). Moreover since *Alu* transcription continues till the Pol III machinery encounters a termination signal, and could thus potentially extend up to position +500 with respect to TSS, we imposed that the 3'segment value is no more than half of the *Alu* body value, to exclude as much as possible noise RNA signal while including downstream extended, genuine *Alu* transcripts. Importantly, only *Alu* transcripts that passed this filter in both replicates were considered to represent autonomously expressed *Alu* loci (as such, they will be often referred to in the text as "expression-positive").

Expression-positive *Alu*s were categorized into the following categories based on their localization: intergenic (that are not hosted in any annotated protein-coding or lincRNA gene; this group also includes *Alu*s that map to introns or exons of annotated genes, but do so in an antisense orientation); intronic Refseq (sense-oriented *Alu*s hosted within introns of RefSeq gene collection); intronic lincRNA (sense-oriented *Alu*s hosted within introns of the lincRNA transcripts annotated in UCSC); 5'UTR and 3'UTR (sense-oriented *Alu*s within UTRs of RefSeq genes). *Alu*s that were not fully contained in any of these genomic locations (e.g. between an exon and an intron) were categorized as "other". Complete lists of these *Alu*s are reported in Supplementary Table S1.

It is evident from this Table that *Alu* transcripts were found both in the Poly(A)+ and Poly(A)- datasets. The same Table also contains a non-redundant list of all expressed *Alu*s obtained by merging expression-positive *Alu*s found in the Poly(A)+ and Poly(A)- fractions of all cell lines ("All non-redundant" sheet in Supplementary Table S1).

To support further the identification of unique *Alu* transcripts found in Hela-S3 and K562 cells, we intersected the ChIP-seq peaks of the Pol III machinery components TFIIIC-110, RPC155, BRF1, BRF2, BDP1, derived from ENCODE/Stanford/Yale/USC/Harvard ChIP-seq data (http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeSydhTfbs/) with the expression-positive *Alu* coordinates, extended to 200 bp upstream, of these two cell lines. The lists of Pol III-associated, expression-positive *Alu*s are reported in Supplementary Table S2.

To identify other transcription factors associated to expression-positive *Alu* elements, we intersected, for each cell line, the 500 bp upstream of the *Alu*s with the coordinates of the transcription factor binding sites from ENCODE ChIP-seq (http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeRegTfbsClustered/ wgEncodeRegTfbsClusteredWithCellsV3.bed.gz). Lists of TF-*Alu* interactions are reported in Supplementary Table S3.

Improved bioinformatic pipeline: *SINEsFind*

This new improved bioinformatic pipeline apply the same concepts seen for the previous one (i.e. filters aimed at excluding SINEs passenger of longer transcripts)

but using an in house developed Python script which allows more control on the data flow and filtering options.

The script first calculates stranded coverage for the whole genome using uniquely mapped reads (NH:i:1) from bam files. Then, for each annotated SINE in the GTF file, it verifies if its maximum coverage peak is greater than the average background noise coverage value in the genome, defined with the –p (--peak) option.

For each SINE element whose peak coverage value was greater than the background, the script first determinates the coordinates of the putative Pol III SINE transcript in order to subsequently define those of the flanking regions and apply the Flanking Region Filter (Supplementary Figure S2).

Since the annotated SINE elements are not always full-length but often represent fragments of the corresponding SINE type, the script perform, for each element, a global alignment using Needleman-Wunsch algorithm between the annotated element sequence and its corresponding full-length consensus sequence to identify the precise element position inside it (i.e. the start of the alignment). Given the alignment coordinates the script calculates the start-end coordinates onto the genome of the corresponding expected full-length SINE defining it as 'central body'. Because known Pol III transcript maximum length is around 500 nt, and SINE elements length (Alus and MIRs) range from 221 to ~300 nt, we defined the 'right arm' starting at the end of the consensus sequence of the element and extending downstream for 200 nt, The 'central body' and 'right arm' regions together represent the maximum span size of a genuine Pol III SINE transcript.

The script then defines as 'left arm' the region starting 100 nt upstream the start coordinate of the 'central arm' and ends 20 nt before it, to include the possibility that in some cases transcription might start slightly upstream of the predicted TSS, usually located ~12-15 nt upstream of the A-box. It also defines an 'out region' that extends 100 nt after the 'right arm'.

To identify SINE transcripts produced by autonomous (most likely Pol III-dependent) transcription, we imposed that the maximum peak coverage value for the 'left arm' is less than the background one (defined with –p) and less than 1/5 of the maximum peak coverage value of the 'central body'. Moreover, since transcription coverage tends to decrease by the end of the transcript we wanted the coverage area of the 'right arm' (as sum of the coverage of each nucleotide in the region) to be less than the 'central body' where most of the transcription should

occur. Finally we imposed the maximum peak coverage value for the 'out' region to be less than the background value to exclude as much as possible noise signal from Pol II transcripts.

MIR ENCODE: Datasets

For MIR RNA identification we used aligned reads (bam files) from datasets described in 7.1.1. To streamline the analyses, and since MIRs do not have A-rich 3' tails, we merged the PolyA+ and PolyA- bam files in each cell line for each replicate thus obtaining 14 bam files (2 replicates for each cell line).

The dataset of annotated MIR was downloaded from UCSC table browser Repeatmasker track of the human genome (GRCh37/hg19). We used MIR elements annotated in intergenic regions, MIR annotated within RefSeq, Ensembl and lincRNA genes and MIR fully contained in introns of these genes in sense orientation and not overlapping with any exon.

Bioinfoinformatic pipeline for individually expressed MIR identification.

We supplied the aligned reads (bam files) along with the annotated MIR in GTF format to the in-house developed Python script (par. 7.2).

We set the -p threshold option in the script to 10 (background peak coverage value), based on the following reasoning. We considered the purely hypothetical case in which all the uniquely mapped reads on hg19 (ranging from 204861066 to 302646644 in the different cell lines) were uniformly distributed onto both strands of the human genome. We then calculated what would be, in this case, the average number of reads for a MIR ~250 bp in length. By dividing the number of uniquely mapped by 6 billion (the stranded human genome size) and multiplying by 250 (the average MIR length) we obtained an average read count ranging from 8.5 and 12.6. We thus imposed to have at least 10 reads overlapping onto each other inside a MIR element.

Importantly, only *MIR* transcripts that passed this filter in both replicates were considered to represent autonomously expressed MIR loci (as such, they will be often referred to in the text as "expression-positive").

To further support the nature of Pol III transcription units of these expression-positive *MIRs* we made use of the Pol3Scan program (5), with less restrictive

parameters (-ca -30 -cb -15 -ct -34), to check for the presence of canonical A and B boxes. Scores closer to 0 represent more likely functional promoters. The program has been run on the genome sequences obtained from the coordinate of their corresponding full-length consensus *MIR*. Of the 188 expression-positive *MIRs* 39 passed the test and coverage profile investigation confirmed their authenticity, along with many other which did not pass the Pol3Scan test. Results are reported in Supplementary Table S5.

Expression-positive MIRs were categorized, based on their localization, in 'intergenic/antisense' (that are either not hosted in any annotated RefSeq gene, Ensembl gene or lincRNA gene or mapped inside those genes but in antisense orientation) and "intronic sense" (that are fully contained within introns of aforementioned genes in sense orientation). Complete lists of these MIRs are reported in Supplementary Table S5. The same Table also contains a non-redundant list of all expressed MIRs obtained by merging expression-positive *MIRs* found in all cell lines ("All unique" sheet in Supplementary Table S5).

To support further the identification of intergenic/antisense unique MIR transcripts found in Hela-S3 and K562 cells, we intersected the ChIP-seq peaks of the Pol III machinery components TFIIIC-110, RPC155, BRF1, BRF2, BDP1, derived from ENCODE/Stanford/Yale/USC/Harvard ChIP-seq data (http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeSydhTfbs/) with the expression-positive MIR coordinates of the corresponding consensus sequence, extended to 200 bp upstream, of these two cell lines. The lists of Pol III-associated, expression-positive MIRs are reported in Supplementary Table S6.

To identify other transcription factors associated to intergenic/antisense expression-positive MIR elements, we intersected, for each cell line, the 500 bp upstream of the corresponding MIRs consensus sequence with the coordinates of the transcription factor binding sites from ENCODE ChIP-seq (http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeRegTfbsClustered/ wgEncodeRegTfbsClusteredWithCellsV3.bed.gz). Lists and statistics of TF-MIR interactions are reported in Supplementary Table S7.

To check if the annotated MIR used in our bioinformatic pipeline were overrepresented in enhancer states of the chromatin (weak and strong) using HMM from ENCODE/Broad for each cell line (excluding HeLa-S3 for which data are not available) we proceeded as follow.

For each chromatin state, we calculated (1) the number of bases (within the submitted intervals) which overlap with that state, and (2) the number of bases in the background (by default, the entire genome) which overlap with that state. For (1) and (2), divide by the number of total bases (in the foreground and background, respectively) to get a "ratio" for each chromatin state (effectively, a fraction of the foreground or background which overlaps with each state). For each chromatin state, divide the foreground ratio by the background ratio, and then take the log (base 10) of that ratio-of-ratios. Basically a log10(Enrichment Score) of 0 means that the chromatin state is in the same fraction in the foreground as the background. A positive value means it was relatively more present in the foreground, and the opposite for a negative value (Supplementary Figure S3).
Again we tested for the enrichment of the expression positive MIRs in the enhancer states performing a Fisher's exact test against all the other annotated MIR used in the pipeline. We found statistical enrichment only in Gm12878 and HepG2 cell lines in strong enhancer state (P-values 1.876e-^{05} and 2.133e-^{05} respectively). (Supplementary Table S8)
To investigate whether expression of MIR elements mapping to Pol II genes in antisense orientation was correlated with the expression of the hosting genes, we calculated normalized reads count for all RefSeq genes (Reflat table from UCSC Table Browser Refseq track of the GRCh37/hg19 assembly) and for our expression positive MIRs inside Pol II genes in antisense orientation using htseq-count tool from HTSeq Python package. We then normalized read counts by the SizeFactor calculated for the RefSeq genes using DESeq2 and calculated correlation using Pearson's coefficient.

To look for variation in the DNA sequence of the sample from which RNA-seq data come from, we made use of samtools to generate Variant Call Format file from aligned RNA-seq reads and bcftools (http://github.com/samtools/bcftools) to reconstruct the consensus DNA sequence with the generated VCF file.

_dl_1500 Ad5 IMR90 infected cells: Bioinformatic pipeline.

Raw RNA-seq reads were aligned to the human genome (GRCH37/hg19) using TopHat (ver. 2.0.11) (1) allowing up to 20 different alignment per read:

```
tophat -p 2 --b2-very-sensitive -o output_dir –library-type
fr-firststrand genome pe1.fastq pe2.fastq
```

For genuine Pol III *Alu* RNA identification we supplied the TopHat-mapped reads (bam files) to the in-house developed *SINEsFind* Python script (see above) along with the annotated *Alu* elements in GTF format. We use only *Alu* elements annotated in intergenic regions and inside RefSeq, Ensembl and lincRNA genes but in antisense orientations (here on referred to as "intergenic/antisense").

We set the -p threshold option in the script to 5 (background peak coverage value), on the basis of the following reasoning. We considered the purely hypothetical case in which all the uniquely mapped reads on hg19 (ranging from 57486640 to 77356608 in the different cell lines) were uniformly distributed onto both strands of the human genome. We then calculated what would be, in this case, the average number of reads for an *Alu* ~300 nt in length. By dividing the number of uniquely mapped by 6 billion (the stranded human genome size) and multiply by 300 (the average *Alu* length) we obtained an average read count ranging from 2.8 and 3.9. We thus imposed to have at least 5 reads overlapping onto each other inside a MIR element.

To further support the authenticity of the expression-positive *Alus* resulting from our bioinformatic pipeline we checked whether they were bound to Pol III TFs. For each expression-positive *Alu* in 24 hr p.i. samples we extended by 200 bp upstream the start coordinate of the corresponding expected full-length *Alu* and intersected its coordinates with the coordinates of the peaks for each ChIP-seq dataset used in the study (ChIP-seq has been performed and data analyzed as described in (6) with 50 nt single end reads). P-values for the association of these TFs were calculated performing a Fisher's exact test against the whole dataset of expected full-length annotated *Alus* used, extended by 200 bp upstream (the coordinates of the expected full-length *Alus* where calculated using the *needle* function in our Python script).

A similar procedure, but performing only Fisher's exact test, has been used to look for enrichment for histone modifications in our expression-positive *Alus* in 24 hr p.i. samples, except that we did not extended by 200 bp the expected full-length Alus coordinates.

The average H3K27ac profiles for mock and dl1500 24 hr p.i. on all *Alus* found to be expressed in the 2 samples, counted once, were plotted using the *sitepro* tools of the CEAS suite (http://liulab.dfci.harvard.edu/CEAS/).

Also for P300 and RB proteins association to mock and *dl*1500 expressed *Alus* 24 hr p.i. we proceeded in the same way but we intersected the peak coordinates for these protein with the coordinate of the 500 bp region upstream the expected full-length *Alus*.

For gene differential expression (DE) we used htseq-count, along with Gencode v.19 annotated genes set, and DESeq Bioconductor package (7). Since we did not have any biological replicate, we first followed the corresponding procedure using the 'blind' method, then we performed the DE analysis considering the 6 hr p.i. samples as biological replicates of the 24 hr p.i. ones to compare fold-change values.
We also used the *sizeFactors* values calculated by DESeq to normalize expression coverage values of our expression-positive *Alus*.

Alu expression profiling in cancer cells

Annotated *Alu* elements have been retrieved from UCSC Table browser Repeatmasker track of the human genome assembly ver. GRChr38/hg19. We used only *Alu* elements annotated in intergenic regions (i.e. outside Refseq, Ensemble and lincRNA genes) since the RNA-seq data used were unstranded and therefore we could not distinguish those annotated inside genes in antisense orientation.
Each RNA-seq dataset was aligned to the human reference genome (GRCh37/hg19) using TopHat ver. 2.0.11:

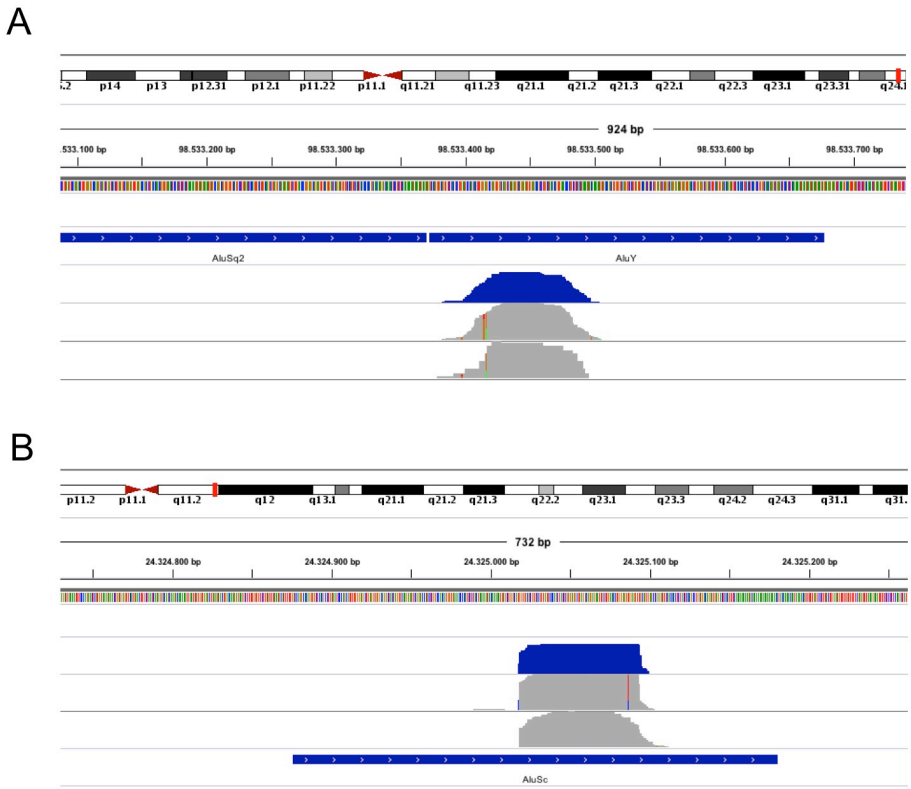
```
$ tophat -p 2 --b2-very-sensitive -o output_dir –library-type fr-unstranded genome pe1.fastq pe2.fastq
```

Each bam file was supplied, along with annotated Alus in GTF format, to our Python script (par. 7.2) using –p 10 as a background coverage value for all the samples. This value has been calculated considering the hypothetical case in which the mapped reads were uniformly distributed onto the "unstranded" human genome. Thus dividing the mapped reads in each sample by 3 billion (the unstranded genome size) and multiplying by 75 (the RNA-seq reads length) we obtained from 2

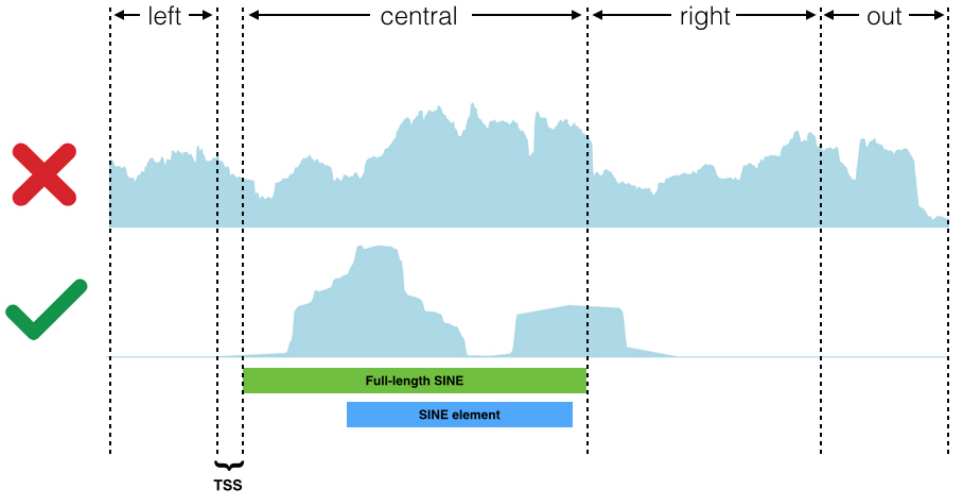
to 9 among all the 36 samples. We then set the background coverage threshold based on the highest value among the samples.

To calculate *Alu* expression levels for a given patient in either condition (tumor and non-tumor), we considered all the Alus found in both tumor and non-tumor samples, counted once, and calculated the expression coverage of the central region for each expected full-length *Alu* in both samples using our *SINEsFind* Python script with *–p 0* and without applying the flanking region filter (because the supplied *Alu* list was already filtered) We did this because of the possibility that an expression-positive *Alu* found in one condition could have been filtered out in the other by our script due to the lower coverage.
To normalize expression values we used the Total Count method. For each sample, its normalization factor is given by the total number of mapped reads in the sample divided by the total number of reads mapped in all the samples (i.e. 72) divided by the number of samples. Thus each expression value in a sample is normalized dividing it by its normalization factor.

Supplementary Figures



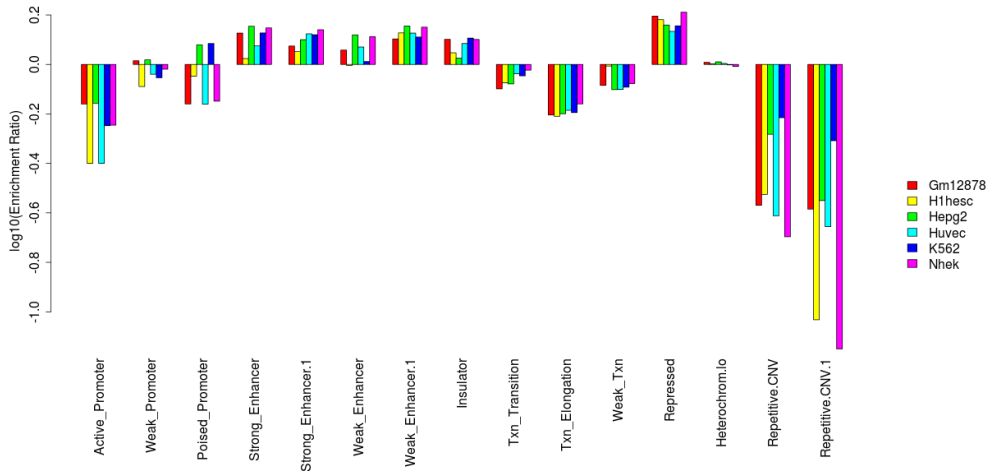
Supplementary Figure S1. (A) Comparison of base-resolution expression profiles observed for AluY(chr10:98533372-98533677) (see Figure 3B of the manuscript) in ENCODE HeLa-S3 polyA+ rep 1, using the following alignment strategies (from top to bottom): TopHat unique alignment (profile generated from bigwig file); TopHat with up to 20 alignments; STAR unique alignment (the last two profiles were generated from bam files).
(B) Comparison of base-resolution expression profiles observed for AluSc(chr14:24324875-24325180) (see Figure 3C of the manuscript) in ENCODE K562 polyA- rep 1, using the following alignment strategies (from top to bottom): TopHat unique alignment (profile generated from bigwig file); TopHat with up to 20 alignments; STAR unique alignment (the last two profiles were generated from bam files).



Supplementary Figure S2. Flanking Region Filter.
To be considered as expression-positive each SINE has to meet the following conditions:

- Peak coverage *left* < background signal AND < 1/5 peak coverage *central*
- Coverage *right* < coverage *central*
- Peak coverage *out* < background signal

For each annotated element having a peak coverage over the background signal, a global alignment with its full-length consensus sequence is performed using Needleman-Wunsch algorithm. The start and end coordinates of the alignment are used to define the *central, left, right* and *out* regions (the latter 3 respectively of 80, 200 and 100 nt). If the peak coverage in the *left* region is under the background value and is < 1/5 of the peak coverage in the *central* region, the coverage of the *right* region is less than the coverage of the *central* region and the peak coverage of the *out* region is under the background value, then the annotated SINE is considered as expression-positive



Supplementary Figure S3. Chromatin state enrichment of annotated MIR used

Supplementary Tables and *SINEsFind*

All the Supplementary Tables and *SINEsFind* Python script can be found at:

https://drive.google.com/folderview?id=0Bxojxm5rb6vScnJjY1QwSWpiRVk&usp=sharing

References

1. Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome biology. 2013;14(4):R36.
2. Treangen TJ, Salzberg SL. Repetitive DNA and next-generation sequencing: computational challenges and solutions. Nature reviews Genetics. 2012;13(1):36-46.
3. Anders S, Pyl TP, Huber W. HTSeq — A Python framework to work with high-throughput sequencing data. . biorXiv preprint. 2014;doi: 10.1101/002824.
4. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009;25(16):2078-9.
5. Percudani R, Pavesi A, Ottonello S. Transfer RNA gene redundancy and translational selection in Saccharomyces cerevisiae. Journal of molecular biology. 1997;268(2):322-30.
6. Ferrari R, Gou D, Jawdekar G, Johnson SA, Nava M, Su T, et al. Adenovirus small E1A employs the lysine acetylases p300/CBP and tumor suppressor Rb to repress select host genes and promote productive virus infection. Cell host & microbe. 2014;16(5):663-76.
7. Anders S, McCarthy DJ, Chen Y, Okoniewski M, Smyth GK, Huber W, et al. Count-based differential expression analysis of RNA sequencing data using R and Bioconductor. Nature protocols. 2013;8(9):1765-86.