An atlas of polymerase III-transcribed *Alu* elements across human cell types and tissues

Xiao-Ou Zhang[1], Thomas R Gingeras[2], Zhiping Weng[1,3,#]

[1]Program in Bioinformatics and Integrative Biology, University of Massachusetts Medical School, Worcester, MA 01605, USA, [2]Functional Genomics, Cold Spring Harbor Laboratory, Cold Spring Harbor, NY 11724, USA, [3]Department of Biochemistry and Molecular Pharmacology, University of Massachusetts Medical School, Worcester, MA 01605, USA

[#]To whom correspondence should be addressed. E-mail: zhiping.weng@umassmed.edu (Z.W.)

Supplemental Material

Supplemental Results and Discussion

Assessment of quality and reproducibility of RAMPAGE data

We assessed the quality of these RAMPAGE data in several ways. First, these RAMPAGE data showed high reproducibility between biological replicates (Supplemental Fig. S1A; Pearson correlation coefficient $R = 0.79 \pm 0.16$ between two biological replicates in 70 replicated biosamples). Second, ~50% of RAMPAGE reads were within 5 nt of annotated TSSs (Supplemental Fig. S1B), while substantially weaker enrichment at annotated TSSs was observed for CAGE reads and no enrichment for RNA-seq reads (Supplemental Fig. S1C). Third, the quantification of gene expression by RAMPAGE significantly correlated with the quantification by RNA-seq (Spearman's rank correlation $\rho = \sim 0.58$; Supplemental Fig. S1D). A similar level of correlation was reported between CAGE and RNA-seq (Kawaji et al. 2014).

Convinced of the high quality of the RAMPAGE data, we built a computational pipeline to call RAMPAGE peaks, which corresponded to individual expressed transcripts. Roughly 96% of RAMPAGE peaks could be assigned to GENCODE-annotated genes or *de novo* assembled transcripts using RNA-seq from the same biosample (Supplemental Fig. S2A). In K562 and GM12878 specifically, roughly 80% of the assigned peaks were within ± 50 nt of the TSSs of GENCODE-annotated genes or *de novo* assembled transcripts (Supplemental Fig. S2A), and some of the remaining 20% peaks might be novel TSSs identified only by RAMPAGE. Indeed, many of these RAMPAGE-only peaks were supported by CAGE data (Abugessaisa et al. 2016) in the same cell type (Supplemental Fig. S2A, pie charts).

Computational validation that the expressed *Alu* elements were Pol III-transcribed

To investigate whether the expressed *Alu* elements identified by our RAMPAGE pipeline resulted from Pol III transcription, we compared the ChIP-seq profiles of POLR3A (the main subunit of Pol III) and POLR2A (the main subunit of Pol II) at these expressed *Alu* elements (Fig. 1C; Supplemental Fig. S3A). Like expressed tRNAs (Moqtaderi et al. 2010; Raha et al. 2010), near 80% of the expressed *Alu* elements

showed high POLR3A signals but lower POLR2A signals around their TSSs, while the opposite was observed for Pol II-transcribed genes. In sharp contrast, randomly sampled, unexpressed *Alu* elements showed almost no signal for POLR3A or POLR2A. Next, we took a closer look at the ChIP-seq data of other components of the Pol III machinery in K562 cells. POLR3A covered the entire body of expressed *Alu* elements, TFIIIC bound the internal Pol III promoter (A-box and B-box), while BDP1 (a major component of TFIIIB) was enriched in the upstream region of *Alu* elements expressed specifically in K562 cells but not in GM12878 cells or other samples (Supplemental Fig. S3B). Consistent with *Alu* elements' being type 2 Pol III targets (Moqtaderi et al. 2010; Oler et al. 2010), much higher ChIP-seq signals of BRF1 than BRF2 were observed in K562 cells at the 5′-end of *Alu* elements expressed in these cells (Supplemental Fig. S3C). ChIP-seq data were available for the TATA-box binding protein TBP (a component of TFIIIB) in both K562 and GM12878 cells, and we observed strong enrichments of TBP binding at the *Alu* elements that were expressed specifically in the corresponding cell types (Supplemental Fig. S3D). These results support we mostly identified PoI III-transcribed primary *Alu* transcripts.

We further asked whether our pipeline could distinguish Pol III-transcribed *Alu* elements from *Alu* elements that were embedded in Pol II-transcribed genes. There is an enrichment of expressed *Alu* elements in the introns of Pol II-transcribed genes. Among the 1.2 M copies of annotated *Alu* elements in the human genome, 24.8% of them reside in Pol II transcribed genes in the sense orientation. In sharp contrast, 78.1% (2,956 out of 3,784) robustly expressed and 76.2% (13,143 out of 17,249) expressed *Alu* elements resided in Pol II transcribed genes in the sense orientation, corresponding to 3.14- and 3.07-fold enrichments, respectively (Chi-squared test p-value $< 2.2\times10^{-16}$ for both). The vast majority of—84.7% of robustly expressed and 93.9% of expressed—sense, genic *Alu* elements were in introns and the remaining in untranslated regions (UTRs), with only one robustly expressed and four expressed sense *Alu* elements overlapping the coding regions of their host Pol II genes. To investigate whether these expressed sense genic *Alu* elements were indeed Pol III-transcribed, and not components of Pol II transcribed genes, we examined Pol III ChIP-seq and RNA-seq data detailed as follows.

Pol III ChIP-seq data revealed that like expressed antisense and intergenic *Alu* elements, expressed sense *Alu* elements showed strong enrichments for POLR3A and TFIIIC binding in K562 cells (Supplemental Fig. S4A), and the enrichments for sense *Alu* elements were statistically indistinguishable from those of intergenic and antisense *Alu* elements (Student's *t*-test *p*-value = 0.52 for POLR3A and 0.59 for TFIIIC).

For embedded *Alu* elements, there should not be a large difference in RNA-seq read density between the *Alu* body and the upstream flanking region (Conti et al. 2015), while we would expect to see a difference for Pol III-transcribed *Alu* elements. RAMPAGE peaks could pinpoint the TSSs of expressed *Alu* elements (Supplemental Fig. S3E), so we directly compared the RNA-seq read density at the ± 100 bp window centered on RAMPAGE peaks. Indeed, we observed lower read densities in the upstream regions of RAMPAGE peaks than the downstream regions for robustly expressed intergenic and intronic *Alu* elements (Fig. 1D shows all such elements, and Supplemental Fig. S4B shows the breakdown into intergenic, sense intronic, and antisense intronic groups), while such a gradient of RNA-seq read density was absent for *Alu* elements located in 3′-UTRs due to the interference of Pol II transcripts (Supplemental Fig. S4B lower right panel). Although the expression levels of the host genes of expressed sense *Alu* elements were significantly higher than those of expressed antisense *Alu* elements (Supplemental Fig. S4C; measured by RNA-seq), the expression levels of their resident sense and antisense *Alu* elements were at similar levels (Supplemental Fig. S4D; measured by RAMPAGE). Thus, both Pol III RNA-seq and RNA-seq data support that the expressed elements identified by our approach were preferentially transcribed by Pol III.

Comparison with previous approaches for identifying expressed *Alu* elements

We compared the expressed *Alu* elements identified by our RAMPAGE pipeline with the expressed *Alu* elements identified in K562 cells by two prior studies, which used RNA-seq (Conti et al. 2015) and Pol III ChIP-seq (Moqtaderi et al. 2010), respectively. To overcome the repetitiveness of *Alu* elements, the

RNA-seq study (Conti et al. 2015) used only uniquely mapping reads; nevertheless, their approach still favored less repetitive *Alu* elements. Furthermore, it is difficult to distinguish primary *Alu* transcripts from embedded *Alu* transcripts relying only on RNA-seq reads. The ChIP-seq studies (Moqtaderi et al. 2010; Oler et al. 2010) identified *Alu* elements bound by POLR3A and/or TFIIIC (using only uniquely mapping reads) but did not assess their transcription.

The overlaps among the three lists of expressed *Alu* elements identified by RAMPAGE (this study; N = 248), RNA-seq (Conti et al. 2015) (N = 154), and POLR3A and TFIIIC ChIP-seq (Moqtaderi et al. 2010) (N = 1,593) are summarized in Fig. 1E. The much longer ChIP-seq list suggests that most of the *Alu* elements bound by POLR3A and TFIIIC were not transcribed. The largest overlap was between the RAMPAGE and ChIP-seq lists—72 (29%) of the 248 expressed *Alu* elements by RAMPAGE had POLR3A and TFIIIC ChIP-seq peaks. In comparison, 29 (19%) of the 154 expressed *Alu* elements by RNA-seq had POLR3A/TFIIIC ChIP-seq peaks. Although only 21 *Alu* elements were deemed expressed by both RAMPAGE and RNA-seq, the vast majority (17; 81%) of these had POLR3A/TFIIIC ChIP-seq peaks. The better agreement between RAMPAGE and POLR3A/TFIIIC ChIP-seq than between RNA-seq and ChIP-seq was further observed by directly examining the ChIP-seq signals. More than 80% of the *Alu* elements uniquely identified by RAMPAGE (N = 227) but less than 30% of the *Alu* elements uniquely defined by RNA-seq (N = 133) showed evident ChIP-seq signals (Fig. 1F), with median POLR3A signals 4.45 vs. 0.80 (p-value = 5.71×10^{-10}) and median TFIIIC signals 5.91 vs. 1.46 (p-value = 6.76×10^{-9}) for the two groups of *Alu* elements (Supplemental Fig. S3F). Even the subset of the *Alu* elements identified by RAMPAGE only but did not result in the calling of ChIP-seq peaks (N = 112) had significantly higher POLR3A/TFIIIC ChIP-seq signals than the subset identified by RNA-seq only (N = 107; Fig. 1G). In summary, we established a method of using RAMPAGE data to identify expressed *Alu* elements genome-wide with improved sensitivity and specificity than earlier approaches.

It remains challenging to quantify the expression levels of highly repetitive *Alu* elements. With the ability to capture precisely the 5′-ends and connect them to the downstream unique sequences, the RAMPAGE assay enabled us to delineate individual *Alu* elements and quantifying those elements that are transcribed. The expressed *Alu* elements that we identified with RAMPAGE data had highly significant, albeit modest, overlap with the expressed *Alu* elements identified by RNA-seq or Pol III ChIP-seq data. Some of the differences are undoubtedly due to the biology, for example, Pol III binding does not necessarily lead to transcription, others could be due to the differences in the assays. The transcripts captured by RAMPAGE are more enriched in 5′-capped RNAs than uncapped, 5′-triphosphorylated RNAs, with *Alu* elements belonging to the latter. Thus, RAMPAGE might have missed some expressed, Pol III-transcribed *Alu* elements. Furthermore, our pipeline may have filtered out some expressed *Alu* elements with low read entropy. On the other hand, RNA-seq based approaches would tend to filter out *Alu* elements with non-uniform read profiles, e.g., many *Alu* elements have few reads at their 5′-halves and they may be filtered out. This issue does not affect our RAMPAGE pipeline because RAMPAGE reads tend to link the 5′-end reads with the downstream locations (Supplemental Fig. S2D). Due to the difference between different assays, inconsistent expression patterns could be observed for some expressed *Alu* elements, especially the intronic ones (Supplemental Fig. S2E). Thus, multiple types of data should be considered when individual *Alu* elements are chosen for experimental testing.

Effect of DICER1 on *Alu* expression

Alu elements are usually transcribed at low levels and accumulation of *Alu* RNAs can cause toxicity and lead to human pathology. DICER1 cleaves *Alu* RNAs, and DICER1 deficiency in the retinal pigmented epithelium can result in accumulation of cytotoxic *Alu* RNAs, which in turn activates innate immune responses, leading to diseases such as age-related macular degeneration (Kaneko et al. 2011; Tarallo et al. 2012). To test whether the global level of *Alu* transcripts was correlated with DICER1 level, we divided the 104 tissue biosamples with both RAMPAGE and RNA-seq data into two equal-sized subsets by the expression level of DICER1. Indeed, the high-DICER1 biosamples had significantly lower global *Alu*

RNA levels than low-DICER1 samples (Wilcoxon rank-sum test *p*-value = 0.03), supporting the effect of DICER1 on *Alu* levels across different cell types.

Specific expression pattern of *Alu* elements in testis

We asked that testis exhibited distinct expression pattern of *Alu* elements from somatic tissues. Among 24 human tissues each derived from at least two individuals, testis was ranked as one of top tissues in the number of expressed *Alu* loci, and significantly higher than 15 tissues but not significantly different from other 8 tissues (Supplemental Fig. S5D), which is consistent with the previous observation of decreased methylation levels at *Alu* loci in testis (Hellmann-Blumberg et al. 1993). However, the expression levels of *Alu* elements in testis are lower than the levels in other tissues (Supplemental Fig. S5E), which may be due to piRNA-mediated silencing in testis (Ha et al. 2014; Williams et al. 2015; Gainetdinov et al. 2017).

Binding of cell type-specific master TFs at expressed *Alu* loci

Some TFs with enrichments of both ChIP-seq signals and sequence motifs at expressed *Alu* loci were only enriched in one cell type but not in the other, such as GATA1::TAL1 in K562 cells and PAX5 in GM12878 cells (Fig. 6A vs. Supplemental Fig. S9A). Often acting in a complex with TAL1, GATA1 is a master regulator of erythropoiesis (Kassouf et al. 2010; Wakabayashi et al. 2016). It is essential for the proliferation of K562 erythroleukemia cells in a dose-dependent manner, and recently a proliferation screen with CRISPR-Cas9 was performed on the *GATA1* locus in K562 cells (Fulco et al. 2016). In contrast, PAX5 is exclusively expressed in the B-lymphoid lineage of the hematopoietic system and is essential for B-lineage commitment (Nutt et al. 1999). These TFs were ranked as the top master transcription factors in the respective cell types based on super-enhancer analysis—TAL1 and GATA1 were ranked the 7th and 10th among 53 K562 master TFs, and PAX5 was ranked the first among 21 GM12878 master TFs (Hnisz et al. 2013). Thus, the binding of master TFs at these cell type-specific *Alu* loci might influence lineage-specific transcriptional programs in corresponding cell types.

Motif enrichment analysis on the *Alu* elements expressed in different tissues also revealed some master transcription factors known to be involved in the biological processes that largely define the identities of the respective tissues (Supplemental Fig. S8C; Supplemental Table S6): TCF4 (Flora et al. 2007; Chen et al. 2016) and MEF2D (Yang et al. 2009) in the brain, TBX5 (Ieda et al. 2010; Qian et al. 2012; Song et al. 2012) and MEF2A (Naya et al. 2002; Schlesinger et al. 2011) in the heart, HNF1A (Kuo et al. 1992; Duncan et al. 1998; Odom et al. 2006), CREB1 (Cereghini 1996; Montminy et al. 2004; Odom et al. 2006), and USF1 (Iynedjian 1998; Pajukanta et al. 2004; Odom et al. 2006) in the liver, TBX5 (Arora et al. 2012), TBX2 (Ludtke et al. 2013), and CEBPA (Martis et al. 2006) in the lung, and IRF1 (Taki et al. 1997) and FLI1 (Suzuki et al. 2013) in the spleen.

Supplemental Methods

RAMPAGE and RNA-seq data

As part of the ENCODE Consortium, we (T. Gingeras's data production center) generated RAMPAGE data in 155 biosamples (Supplemental Table S1). The data, metadata, and protocol of these experiments are all available at the ENCODE Portal (https://www.encodeproject.org/). We also generated RNA-seq data for all of these 155 biosamples (Supplemental Table S1), and the data, metadata, and protocol of these experiments are also available at the ENCODE Portal. These data were processed using the ENCODE uniform processing pipelines for the respective data type (https://www.encodeproject.org/pipelines/). RNA-seq alignment (BAM files) and gene expression quantification are part of the outputs of the pipeline and available at the ENCODE Portal. We used these files for downstream analysis.

ChIP-seq and CAGE data processing

Alignment and quantification of ChIP-seq were performed by the ENCODE Data Coordinating Center using the ENCODE uniform processing pipeline of ChIP-seq data (https://www.encodeproject.org/pipelines/). Normalized ChIP-seq signals (fold change of ChIP over input) were downloaded from the ENCODE Portal for downstream analysis (Supplemental Table S1). Read alignment and peak calling of CAGE data were carried out by the FANTOM Consortium (The FANTOM Consortium and the RIKEN PMI and CLST (DGT) 2014), and alignment files (BAM) and CAGE peaks were downloaded from the FANTOM5 SSTAR database (Abugessaisa et al. 2016) (http://fantom.gsc.riken.jp/5/sstar). Detailed information on all ChIP-seq and CAGE datasets is included in Supplemental Table S1.

Peak calling for RAMPAGE data and identification of primary *Alu* transcripts

We built a computational pipeline to identify RAMPAGE peaks, each annotating an expressed transcript. The BAM files of 155 RAMPAGE datasets (Supplemental Table S1) were downloaded from the ENCODE Portal (https://www.encodeproject.org/); these files contained sequencing reads aligned to the GRCh37/hg19 reference human genome using STAR (Dobin et al. 2013) based on the ENCODE RAMPAGE Processing Pipeline (https://www.encodeproject.org/rampage/). Among the 155 datasets, 70 contained two biological replicates (hence two BAM files) while the remaining 85 had only one replicate (due to the limitation of tissue samples). As illustrated in Fig. 1A, we identified all properly aligned read pairs (R1 and R2 denote the paired reads) with uniquely mapped R2 reads from each BAM file and collapsed the read pairs with the same alignment coordinates and the identical 15-bp barcode at the 5′-end of the R2 reads, as these were presumably PCR duplicates (Batut and Gingeras 2013; Fu et al. 2018). Reads from biological replicates were pooled together after the PCR duplicate removal. The 5′-most base of each aligned read pairs, which corresponded to the TSSs of the expressed transcripts, were then clustered into RAMPAGE peaks using F-seq (Boyle et al. 2008) with parameter settings of feature length = 30 and fragment size = 0. To eliminate lengthy tails, we resized the RAMPAGE peaks to the length that included 95% of the reads in each peak (Ni et al. 2010). The entropy (E) of a peak was calculated based on the coordinates of R2 reads in this peak. The effective length (L) of a peak was defined as the shortest distance from the 5′-end that contained 75% of read pairs in the peak.

We then compared our RAMPAGE peaks with annotated genes from GENCODE (V19) and annotated *Alu* elements from the UCSC Genome Browser (hg19 rmsk.txt downloaded from UCSC). To account for alternative promoters, we extended annotated *Alu* elements by padding 50 nts to their annotated 5′-ends when assigning RAMPAGE peaks to specific *Alu* elements. We supplemented GENCODE genes with *de novo* assembled transcripts (StringTie v1.3.3b (Pertea et al. 2015) with default parameters) using RNA-seq data in the corresponding biosample. As shown in Supplemental Fig. S2C, the RAMPAGE peaks that overlapped the TSSs of annotated genes had much higher entropies than the peaks that did not, and an entropy of 2.5 separated the two groups of peaks. Furthermore, the effective lengths of most RAMPAGE

peaks that overlapped annotated *Alu* elements clustered around 280 nts, the consensus length of full-length *Alu* elements, and few peaks exceeded 1 k nucleotides (Supplemental Fig. S2D). We also tested our pipeline with only unspliced RAMPAGE reads, and 99% (247 out of 248) of K562 and 100% (140 out of 140) GM12878 expressed *Alu* elements could be discovered, suggesting that the effective length cutoff could effectively remove spliced transcripts. Thus, we only retained those RAMPAGE peaks with $E \geq 2.5$ and $L \leq 1$ k nucleotides for further analysis. To minimize the contamination from non-*Alu* transcripts, RAMPAGE peaks with read pairs covering over 50% of an annotated *Alu* body were deemed candidate *Alu* RAMPAGE peaks, and the corresponding *Alu* elements were deemed expressed. If one RAMPAGE peak overlapped two adjacent *Alu* elements, this peak was assigned to the *Alu* with the shortest distance to its 5′-end. If one expressed *Alu* element had multiple RAMPAGE peaks, the peak with the highest expression level (measured by RPM, reads per million mapped reads) was retained. As we wanted to identify the primary *Alu* transcripts transcribed by Pol III, we further eliminated the RAMPAGE peaks that overlapped the ± 250 bp windows centered on the TSSs of any annotated genes.

To assess the impact of human reference genome version on the identification of expressed *Alu* elements, we reapplied our computational pipeline to the RAMPAGE data by aligning to GRCh38 and overlapped with the expressed *Alu* elements identified in hg19 (Supplemental Table S2). Ninety-seven percent (241 out of 248) of K562 and 96% (135 out of 140) of GM12878 expressed *Alu* elements identified in hg19 were also detected in GRCh38, and only four and six expressed *Alu* elements in K562 and GM12878 cells were newly annotated in GRCh38, respectively. Among the 155 biosamples, 99% and 84% of biosamples had discovery rates higher than 80% and 90%, respectively.

Comparison of RAMPAGE peaks with Pol III ChIP-seq data and RNA-seq data

ChIP-seq peaks of POLR3A and TFIIIC in K562 cells were obtained from (Moqtaderi et al. 2010) and overlapped with *Alu* annotations (hg19 rmsk.txt downloaded from UCSC) to identify Pol III bound *Alu* elements. The list of expressed *Alu* elements in K562 defined using RNA-seq was downloaded from

(Conti et al. 2015), and the binding profile of POLR3A and TFIIIC ChIP-seq in K562 were compared with expressed *Alu* elements identified by RAMPAGE, RNA-seq, or both techniques (Fig. 1E-G; Supplemental Fig. S3F).

Comparison of expressed *Alu* elements with unexpressed *Alu* elements, tRNAs, and Pol II-transcribed genes

Unexpressed *Alu* elements in K562 cells were randomly sampled from genomic *Alu* elements that did not have any expression signals in either of the two RAMPAGE libraries that corresponding to the biological replicates. The list of expressed tRNAs in K562 cells was obtained from (Raha et al. 2010). Pol II-transcribed genes were randomly sampled from genes with high expression (RPM $\geq$ 10) according to K562 RAMPAGE data. Since tRNAs generally have extremely high Pol III binding signals, *Alu* elements near tRNAs ($\leq$ 1 kb) were removed for the Pol III binding analyses.

Anatomical classification of tissues

The UBERON anatomy ontology term (Mungall et al. 2012) of each tissue (Supplemental Table S3) was extracted from ENCODE Biosample database (https://www.encodeproject.org/). Ontology terms were manually classified based on indirect and direct 'is_a' or 'part_of' relationships to construct mutually exclusive organ groups for the tissue samples (Fig. 2C; Supplemental Fig. S5C).

Evolutionary analysis

Using the liftOver tool with whole-genome alignment-chain files downloaded from the UCSC, we mapped 17,249 expressed human *Alu* elements to four other primate genomes: chimpanzee (panTro5), gorilla (gorGor3), orangutan (ponAbe2) and rhesus (rheMac8). We used the primate phylogeny reported previously: ((((Human, Chimpanzee), Gorilla), Orangutan), Rhesus) (Locke et al. 2011). *Alu* elements that could not be lifted over to another primate genome were considered missing in that genome. The origin of an *Alu* element was defined as the last common ancestor that descended to all species with its

orthologs (Fig. 3A; Supplemental Fig. S6C; Supplemental Table S4). For example, if a human *Alu* element was in the gorilla genome but not in the other three primate genomes, this *Alu* belonged to B2 (Fig. 3A). Sequence divergence of each *Alu* element from the consensus sequence was calculated with RepeatMasker and reported as the total number of mismatches per 100 nts (Fig. 3C; Supplemental Fig. S6B).

Motif analysis

The sequence motifs of A-box and B-box (Supplemental Fig. S6F) were identified using the *de novo* motif discovery algorithm MEME (Bailey et al. 2009) on 5,000 randomly sampled expressed or unexpressed *Alu* sequences separately, and the local enrichment of A-box or B-box in expressed and unexpressed *Alu* elements (Supplemental Fig. S6G) was carried out using CentriMo (Bailey and Machanick 2012) (parameters: --local --minreg 10 --norc). To search for TF binding motifs, we used the AME algorithm (McLeay and Bailey 2010) to identify annotated motifs from the JASPAR CORE vertebrates database (v2018) (Sandelin et al. 2004) that might be enriched in the TSS ± 500 bp region of expressed *Alu* elements in tissues and cell lines, respectively (Supplemental Table S6).

DNA methylation analysis

Sequencing reads of whole-genome bisulfite sequencing datasets in K562 and GM12878 cells were first trimmed using Trim Galore (https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/) with default parameters to remove adapters and low-quality bases. Trimmed reads were aligned to the human reference genome (hg19) with Bismark (Krueger and Andrews 2011) (parameters: --bowtie2 -N 1 -L 28), and the methylation call for every CpG was extracted by Bismark's methylation extractor script. CpGs covered by at least 3 reads were extracted (Tsuji and Weng 2016) and then overlapped with *Alu* annotations. DNA methylation levels of CpGs were compared between expressed *Alu* elements and unexpressed *Alu* elements (Fig. 4B).

Random forest approach to classify expressed *Alu* elements

We used the random forest classifier implemented by the h2o Python package (https://github.com/h2oai/h2o-3) to train models for predicting the transcriptional states of *Alu* elements (robustly expressed or expressed vs. unexpressed). Random forest classifiers were implemented using 500 classification trees sampled at each split and a maximum depth of 10 with 10-fold cross-validation (parameters: ntrees = 500, nfolds = 10, max_depth = 10, balance_classes = True, keep_cross_validation_predictions = True).

To characterize the impact of genomic context and primary sequences on *Alu* transcription (Fig. 3G, H), two random forest classifiers were trained to distinguish robustly expressed or expressed *Alu* elements against 10,000 randomly sampled, unexpressed *Alu* elements. The classifiers chose from 42 input features: (a) distance to the TSS of the nearest Pol II-transcribed gene (gene distance) or the nearest *Alu* element (*Alu* distance); (b) counts of nearby (± 10 kb) Pol II-transcribed genes (gene count) or *Alu* elements (*Alu* count); (c) *Alu* subfamily (Y, S, or J); (d) length of the *Alu* element; (e) the average phastCons sequence conservation score for each *Alu* element, extracted from the UCSC multiple alignments of primate genomes (http://hgdownload.cse.ucsc.edu/goldenpath/hg19/database/phastCons46wayPrimates.txt.gz); (f) GC content and dinucleotide frequencies of the *Alu* element and its 100-bp upstream region; and (g) sequence divergence from the *Alu* consensus sequence.

We also trained random forest classifiers that additionally included epigenetic features (Fig. 4D). These classifiers were trained on balanced classes of *Alu* elements expressed in a specific cell type (K562, GM12878, or PC-3) against 150 randomly sampled *Alu* elements expressed in other cell types. Input features included normalized DNase-seq signal, ChIP-seq signals of H2AZ, H3K27ac, H3K27me3, H3K36me3, H3K4me1, H3K4me2, H3K4me3, H3K79me2, H3K9ac, H3K9me3, and H4K20me1, the

percentage of CpG methylation and related genomic context features (gene distance, gene count, *Alu* distance, *Alu* count, and *Alu* subfamily).

Gene Ontology analysis

Alu elements specifically expressed in brain, heart, liver, lung, or spleen were identified using RAMPAGE datasets that corresponded to the tissues (Supplemental Table S3), and GO term enrichment on proximal genes was performed using GREAT (McLean et al. 2010) (http://great.stanford.edu) with the following parameter settings: basal plus extension, proximal 5 kb upstream and 1 kb downstream, plus distal up to 1,000 kb. The top 10 enriched GO terms for each tissue with FDR < 0.05 were plotted in Fig. 5B and listed in Supplemental Table S5.

Overlap with candidate cis-regulatory elements and chromatin states

The ENCODE cell type-agnostic and cell type-specific candidate Cis-Regulatory Elements (cCREs) and the subset of cCREs with enhancer-like signatures (cCREs-ELS) were downloaded from the SCREEN resource (http://screen.encodeproject.org/). Chromatin states were downloaded from the ENCODE ChromHMM annotations (Ernst et al. 2011) for K562 and GM12878 (Fig. 5E).

ChIA-PET data analysis

We downloaded paired-end-tag (PET) clusters of Pol II ChIA-PET data in K562 (Li et al. 2012) and GM12878 (Tang et al. 2015). PET clusters were further filtered to exclude inter-chromosomal interactions. An *Alu* element was considered linked to the promoter of a protein-coding gene if both the *Alu* and the promoter overlapped tags of the same PET cluster. cCREs-ELS *Alu* elements, non-cCREs-ELS *Alu* elements, and *Alu* elements expressed in other biosamples were analyzed and compared (Fig. 5F).

Enrichment of transcription factor binding at expressed *Alu* loci

The ChIP-seq signals of TFs in the TSS ± 500 bp window of expressed *Alu* elements were quantified in K562 and GM12878 cells, respectively. To control for genomic context, 500 randomly sampled *Alu* elements expressed in other biosamples but not in K562 or GM12878 were used as control (Fig. 6; Supplemental Fig. S9).

Overlap with STARR-seq, Sharpr-MPRA, and CRISPR-QTL data

STARR-seq data of primed ESCs and naive ESCs were downloaded from Barakat et al. and active enhancers were defined as RPP (reads per plasmid) ≥ 138 (Barakat et al. 2018). Expressed *Alu* elements in H7 and induced pluripotent stem cell derived from fibroblast of arm were combined as *Alu* elements expressed in ESCs (Supplemental Fig. S10A). Sharpr-MPRA activity scores of 15,720 regions tested in K562 cells were downloaded from (Ernst et al. 2016), and activity scores at the promoter region (TSS ± 50 bp) of expressed and unexpressed *Alu* elements were compared (Supplemental Fig. S10B). The enhancer-gene pairs identified in K562 using CRISPR-QTL were downloaded from (Gasperini et al. 2019) and the *Alu* elements expressed in K562 were overlapped with the enhancers to identify their target genes (Fig. 6B).

References

Abugessaisa I, Shimoji H, Sahin S, Kondo A, Harshbarger J, Lizio M, Hayashizaki Y, Carninci P, Forrest A, Kasukawa T et al. 2016. FANTOM5 transcriptome catalog of cellular states based on Semantic MediaWiki. *Database : the journal of biological databases and curation* **2016**.

Arora R, Metzger RJ, Papaioannou VE. 2012. Multiple roles and interactions of Tbx4 and Tbx5 in development of the respiratory system. *PLoS genetics* **8**: e1002866.

Bailey TL, Boden M, Buske FA, Frith M, Grant CE, Clementi L, Ren J, Li WW, Noble WS. 2009. MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res* **37**: W202-208.

Bailey TL, Machanick P. 2012. Inferring direct DNA binding from ChIP-seq. *Nucleic Acids Res* **40**: e128.

Barakat TS, Halbritter F, Zhang M, Rendeiro AF, Perenthaler E, Bock C, Chambers I. 2018. Functional Dissection of the Enhancer Repertoire in Human Embryonic Stem Cells. *Cell stem cell* **23**: 276-288.e278.

Batut P, Gingeras TR. 2013. RAMPAGE: Promoter Activity Profiling by Paired-End Sequencing of 5′-Complete cDNAs. *Current Protocols in Molecular Biology* **104**: 25B.11.21-25B.11.16.

Boyle AP, Guinney J, Crawford GE, Furey TS. 2008. F-Seq: a feature density estimator for high-throughput sequence tags. *Bioinformatics* **24**: 2537-2538.

Cereghini S. 1996. Liver-enriched transcription factors and hepatocyte differentiation. *FASEB journal : official publication of the Federation of American Societies for Experimental Biology* **10**: 267-282.

Chen T, Wu Q, Zhang Y, Lu T, Yue W, Zhang D. 2016. Tcf4 Controls Neuronal Migration of the Cerebral Cortex through Regulation of Bmp7. *Frontiers in molecular neuroscience* **9**: 94.

Conti A, Carnevali D, Bollati V, Fustinoni S, Pellegrini M, Dieci G. 2015. Identification of RNA polymerase III-transcribed Alu loci by computational screening of RNA-Seq data. *Nucleic Acids Res* **43**: 817-835.

Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. 2013. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**: 15-21.

Duncan SA, Navas MA, Dufort D, Rossant J, Stoffel M. 1998. Regulation of a transcription factor network required for differentiation and metabolism. *Science* **281**: 692-695.

Ernst J, Kheradpour P, Mikkelsen TS, Shoresh N, Ward LD, Epstein CB, Zhang X, Wang L, Issner R, Coyne M et al. 2011. Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature* **473**: 43-49.

Ernst J, Melnikov A, Zhang X, Wang L, Rogov P, Mikkelsen TS, Kellis M. 2016. Genome-scale high-resolution mapping of activating and repressive nucleotides in regulatory regions. *Nature biotechnology* **34**: 1180-1190.

Flora A, Garcia JJ, Thaller C, Zoghbi HY. 2007. The E-protein Tcf4 interacts with Math1 to regulate differentiation of a specific subset of neuronal progenitors. *Proceedings of the National Academy of Sciences of the United States of America* **104**: 15382-15387.

Fu Y, Wu P-H, Beane T, Zamore PD, Weng Z. 2018. Elimination of PCR duplicates in RNA-seq and small RNA-seq using unique molecular identifiers. *BMC Genomics* **19**: 531.

Fulco CP, Munschauer M, Anyoha R, Munson G, Grossman SR, Perez EM, Kane M, Cleary B, Lander ES, Engreitz JM. 2016. Systematic mapping of functional enhancer-promoter connections with CRISPR interference. *Science* **354**: 769-773.

Gainetdinov I, Skvortsova Y, Kondratieva S, Funikov S, Azhikina T. 2017. Two modes of targeting transposable elements by piRNA pathway in human testis. *RNA* **23**: 1614-1625.

Gasperini M, Hill AJ, McFaline-Figueroa JL, Martin B, Kim S, Zhang MD, Jackson D, Leith A, Schreiber J, Noble WS et al. 2019. A Genome-wide Framework for Mapping Gene Regulation via Cellular Genetic Screens. *Cell* **176**: 377-390.e319.

Ha H, Song J, Wang S, Kapusta A, Feschotte C, Chen KC, Xing J. 2014. A comprehensive analysis of piRNAs from adult human testis and their relationship with genes and mobile elements. *BMC Genomics* **15**: 545.

Hellmann-Blumberg U, Hintz MF, Gatewood JM, Schmid CW. 1993. Developmental differences in methylation of human Alu repeats. *Molecular and cellular biology* **13**: 4523.

Hnisz D, Abraham BJ, Lee TI, Lau A, Saint-Andre V, Sigova AA, Hoke HA, Young RA. 2013. Super-enhancers in the control of cell identity and disease. *Cell* **155**: 934-947.

Ieda M, Fu JD, Delgado-Olguin P, Vedantham V, Hayashi Y, Bruneau BG, Srivastava D. 2010. Direct reprogramming of fibroblasts into functional cardiomyocytes by defined factors. *Cell* **142**: 375-386.

Iynedjian PB. 1998. Identification of upstream stimulatory factor as transcriptional activator of the liver promoter of the glucokinase gene. *The Biochemical journal* **333 (Pt 3)**: 705-712.

Kaneko H, Dridi S, Tarallo V, Gelfand BD, Fowler BJ, Cho WG, Kleinman ME, Ponicsan SL, Hauswirth WW, Chiodo VA et al. 2011. DICER1 deficit induces Alu RNA toxicity in age-related macular degeneration. *Nature* **471**: 325-330.

Kassouf MT, Hughes JR, Taylor S, McGowan SJ, Soneji S, Green AL, Vyas P, Porcher C. 2010. Genome-wide identification of TAL1's functional targets: insights into its mechanisms of action in primary erythroid cells. *Genome Res* **20**: 1064-1083.

Kawaji H, Lizio M, Itoh M, Kanamori-Katayama M, Kaiho A, Nishiyori-Sueki H, Shin JW, Kojima-Ishiyama M, Kawano M, Murata M et al. 2014. Comparison of CAGE and RNA-seq transcriptome profiling using clonally amplified and single-molecule next-generation sequencing. *Genome Res* **24**: 708-717.

Krueger F, Andrews SR. 2011. Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. *Bioinformatics* **27**: 1571-1572.

Kuo CJ, Conley PB, Chen L, Sladek FM, Darnell JE, Jr., Crabtree GR. 1992. A transcriptional hierarchy involved in mammalian cell-type specification. *Nature* **355**: 457-461.

Li G, Ruan X, Auerbach RK, Sandhu KS, Zheng M, Wang P, Poh HM, Goh Y, Lim J, Zhang J et al. 2012. Extensive promoter-centered chromatin interactions provide a topological basis for transcription regulation. *Cell* **148**: 84-98.

Locke DP Hillier LW Warren WC Worley KC Nazareth LV Muzny DM Yang S-P Wang Z Chinwalla AT Minx P et al. 2011. Comparative and demographic analysis of orang-utan genomes. *Nature* **469**: 529.

Ludtke TH, Farin HF, Rudat C, Schuster-Gossler K, Petry M, Barnett P, Christoffels VM, Kispert A. 2013. Tbx2 controls lung growth by direct repression of the cell cycle inhibitor genes Cdkn1a and Cdkn1b. *PLoS genetics* **9**: e1003189.

Martis PC, Whitsett JA, Xu Y, Perl AK, Wan H, Ikegami M. 2006. C/EBPalpha is required for lung maturation at birth. *Development (Cambridge, England)* **133**: 1155-1164.

McLean CY, Bristor D, Hiller M, Clarke SL, Schaar BT, Lowe CB, Wenger AM, Bejerano G. 2010. GREAT improves functional interpretation of cis-regulatory regions. *Nature biotechnology* **28**: 495-501.

McLeay RC, Bailey TL. 2010. Motif Enrichment Analysis: a unified framework and an evaluation on ChIP data. *BMC Bioinformatics* **11**: 165.

Montminy M, Koo SH, Zhang X. 2004. The CREB family: key regulators of hepatic metabolism. *Annales d'endocrinologie* **65**: 73-75.

Moqtaderi Z, Wang J, Raha D, White RJ, Snyder M, Weng Z, Struhl K. 2010. Genomic binding profiles of functionally distinct RNA polymerase III transcription complexes in human cells. *Nat Struct Mol Biol* **17**: 635-640.

Mungall CJ, Torniai C, Gkoutos GV, Lewis SE, Haendel MA. 2012. Uberon, an integrative multi-species anatomy ontology. *Genome Biol* **13**: R5.

Naya FJ, Black BL, Wu H, Bassel-Duby R, Richardson JA, Hill JA, Olson EN. 2002. Mitochondrial deficiency and cardiac sudden death in mice lacking the MEF2A transcription factor. *Nature medicine* **8**: 1303-1309.
Ni T, Corcoran DL, Rach EA, Song S, Spana EP, Gao Y, Ohler U, Zhu J. 2010. A paired-end sequencing strategy to map the complex landscape of transcription initiation. *Nat Methods* **7**: 521-527.
Nutt SL, Heavey B, Rolink AG, Busslinger M. 1999. Commitment to the B-lymphoid lineage depends on the transcription factor Pax5. *Nature* **401**: 556-562.
Odom DT, Dowell RD, Jacobsen ES, Nekludova L, Rolfe PA, Danford TW, Gifford DK, Fraenkel E, Bell GI, Young RA. 2006. Core transcriptional regulatory circuitry in human hepatocytes. *Molecular systems biology* **2**: 2006.0017.
Oler AJ, Alla RK, Roberts DN, Wong A, Hollenhorst PC, Chandler KJ, Cassiday PA, Nelson CA, Hagedorn CH, Graves BJ et al. 2010. Human RNA polymerase III transcriptomes and relationships to Pol II promoter chromatin and enhancer-binding factors. *Nat Struct Mol Biol* **17**: 620-628.
Pajukanta P, Lilja HE, Sinsheimer JS, Cantor RM, Lusis AJ, Gentile M, Duan XJ, Soro-Paavonen A, Naukkarinen J, Saarela J et al. 2004. Familial combined hyperlipidemia is associated with upstream transcription factor 1 (USF1). *Nat Genet* **36**: 371-376.
Pertea M, Pertea GM, Antonescu CM, Chang TC, Mendell JT, Salzberg SL. 2015. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nature biotechnology* **33**: 290-295.
Qian L, Huang Y, Spencer CI, Foley A, Vedantham V, Liu L, Conway SJ, Fu J-d, Srivastava D. 2012. In vivo reprogramming of murine cardiac fibroblasts into induced cardiomyocytes. *Nature* **485**: 593.
Raha D, Wang Z, Moqtaderi Z, Wu L, Zhong G, Gerstein M, Struhl K, Snyder M. 2010. Close association of RNA polymerase II and many transcription factors with Pol III genes. *Proceedings of the National Academy of Sciences* **107**: 3639-3644.
Sandelin A, Alkema W, Engstrom P, Wasserman WW, Lenhard B. 2004. JASPAR: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Res* **32**: D91-94.
Schlesinger J, Schueler M, Grunert M, Fischer JJ, Zhang Q, Krueger T, Lange M, Tonjes M, Dunkel I, Sperling SR. 2011. The cardiac transcription network modulated by Gata4, Mef2a, Nkx2.5, Srf, histone modifications, and microRNAs. *PLoS genetics* **7**: e1001313.
Song K, Nam Y-J, Luo X, Qi X, Tan W, Huang GN, Acharya A, Smith CL, Tallquist MD, Neilson EG et al. 2012. Heart repair by reprogramming non-myocytes with cardiac transcription factors. *Nature* **485**: 599.
Suzuki E, Williams S, Sato S, Gilkeson G, Watson DK, Zhang XK. 2013. The transcription factor Fli-1 regulates monocyte, macrophage and dendritic cell development in mice. *Immunology* **139**: 318-327.
Taki S, Sato T, Ogasawara K, Fukuda T, Sato M, Hida S, Suzuki G, Mitsuyama M, Shin EH, Kojima S et al. 1997. Multistage regulation of Th1-type immune responses by the transcription factor IRF-1. *Immunity* **6**: 673-679.
Tang Z, Luo OJ, Li X, Zheng M, Zhu JJ, Szalaj P, Trzaskoma P, Magalska A, Wlodarczyk J, Ruszczycki B et al. 2015. CTCF-Mediated Human 3D Genome Architecture Reveals Chromatin Topology for Transcription. *Cell* **163**: 1611-1627.
Tarallo V, Hirano Y, Gelfand BD, Dridi S, Kerur N, Kim Y, Cho WG, Kaneko H, Fowler BJ, Bogdanovich S et al. 2012. DICER1 loss and Alu RNA induce age-related macular degeneration via the NLRP3 inflammasome and MyD88. *Cell* **149**: 847-859.
The FANTOM Consortium and the RIKEN PMI and CLST (DGT). 2014. A promoter-level mammalian expression atlas. *Nature* **507**: 462-470.
Tsuji J, Weng Z. 2016. Evaluation of preprocessing, mapping and postprocessing algorithms for analyzing whole genome bisulfite sequencing data. *Brief Bioinform* **17**: 938-952.
Wakabayashi A, Ulirsch JC, Ludwig LS, Fiorini C, Yasuda M, Choudhuri A, McDonel P, Zon LI, Sankaran VG. 2016. Insight into GATA1 transcriptional activity through interrogation of cis

elements disrupted in human erythroid disorders. *Proceedings of the National Academy of Sciences of the United States of America* **113**: 4434-4439.
Williams Z, Morozov P, Mihailovic A, Lin C, Puvvula Pavan K, Juranek S, Rosenwaks Z, Tuschl T. 2015. Discovery and Characterization of piRNAs in the Human Fetal Ovary. *Cell Reports* **13**: 854-863.
Yang Q, She H, Gearing M, Colla E, Lee M, Shacka JJ, Mao Z. 2009. Regulation of neuronal survival factor MEF2D by chaperone-mediated autophagy. *Science* **323**: 124-127.