G3 Genes | Genomes | Genetics

INVESTIGATIONS

RNA transcription and degradation of Alu retrotransposons depends on sequence features and evolutionary history

Till Baar[1,*], **Sebastian Dümcke**[†], **Saskia Gressel**[‡], **Björn Schwalb**[‡], **Alexander Dilthey**[§], **Patrick Cramer**[‡] **and Achim Tresch**[1,*,**,††]

[*]Institute of Medical Statistics and Computational Biology, Faculty of Medicine, University of Cologne, 50937 Cologne, Germany, [†]Clemedi AG, Wagistr. 12, 8952 Schlieren, Switzerland, [‡]Department of Molecular Biology, Max Planck Institute for Biophysical Chemistry, 37077 Göttingen, Germany, [§]Institute of Medical Microbiology and Hospital Hygiene, Medical Faculty, Heinrich-Heine-University Düsseldorf, 40225 Düsseldorf, Germany, [**]CECAD, University of Cologne, 50931 Cologne, Germany, [††]Center for Data and Simulation Science, University of Cologne, 50923 Cologne, Germany

ABSTRACT Alu elements are one of the most successful groups of RNA retrotransposons and make up 11% of the human genome with over one million individual loci. They are linked to genetic defects, increases in sequence diversity, and influence transcriptional activity. Still, their RNA metabolism is poorly understood yet. It is even unclear whether Alu elements are mostly transcribed by RNA Polymerase II or III. We have conducted a transcription shutoff experiment by α-amanitin and metabolic RNA labeling by 4-thiouridine combined with RNA fragmentation (TT-seq) and RNA-seq to shed further light on the origin and life cycle of Alu transcripts. We find that Alu RNAs are more stable than previously thought and seem to originate in part from RNA Polymerase II activity, as previous reports suggest. Their expression however seems to be independent of the transcriptional activity of adjacent genes. Furthermore, we have developed a novel statistical test for detecting expression quantitative trait loci in Alu elements that relies on the de Bruijn graph representation of all Alu sequences. It controls for both statistical significance and biological relevance using a tuned k-mer representation, discovering influential sequence features missed by regular motif search. In addition, we discover several point mutations using a generalized linear model, and motifs of interest, which also match transcription factor binding motifs.

KEYWORDS
Alu elements
Alu transcription
RNA labeling
retrotransposons

INTRODUCTION

Alu elements are around 300 bp long RNA retrotransposons (Lander *et al.* 2001, for review see Deininger *et al.* 2011). They are classified as short interspersed nuclear elements (SINEs) and are highly abundant in the genome of higher primates (Gentles *et al.* 2005). Over one million Alu loci are currently annotated in the human genome. This means that 11% of the whole genome consists of just Alu sequences (Lander *et al.* 2001), making them one of the most successful groups of mobile elements.

Manuscript compiled: Tuesday 1st March, 2022
[1]Corresponding authors: Till Baar, Institut für medizinische Statistik und Bioinformatik, Robert-Koch-Straße 10, 50931 Köln, Deutschland | till.baar@uni-koeln.de — Achim Tresch, Institut für medizinische Statistik und Bioinformatik, Robert-Koch-Straße 10, 50931 Köln, Deutschland | achim.tresch@uni-koeln.de

Alu elements were discovered using a restriction endonuclease of *Athrobacter luteus*, which gave them their name (Schmid and Deininger 1975). They are RNA retrotransposons, also called class I transposable elements, and as such are capable of copying themselves into new positions in the genome.

As shown schematically in Figure 1a, once an Alu element is transcribed, its RNA attaches itself to the exit tunnel of the ribosome, using a sequence homolog to the signal recognition particle (SRP), a ribonucleoprotein part of the eukaryotic ribosome (Conti *et al.* 2015; Tisdale and Pellizzoni 2017). As Alu elements are nonautonomous retrotransposons, they lack their own retrotransposase domain. Instead, they rely on autonomous LINE-1 elements, another class of retrotransposon belonging to the long interspersed nuclear elements (LINEs) that cover up to 17% of the genome (Boeke 1997; Dewannieux *et al.* 2003). Thus, the Alu RNA lies in wait at the ribosome's exit tunnel until a LINE-1 element is trans-

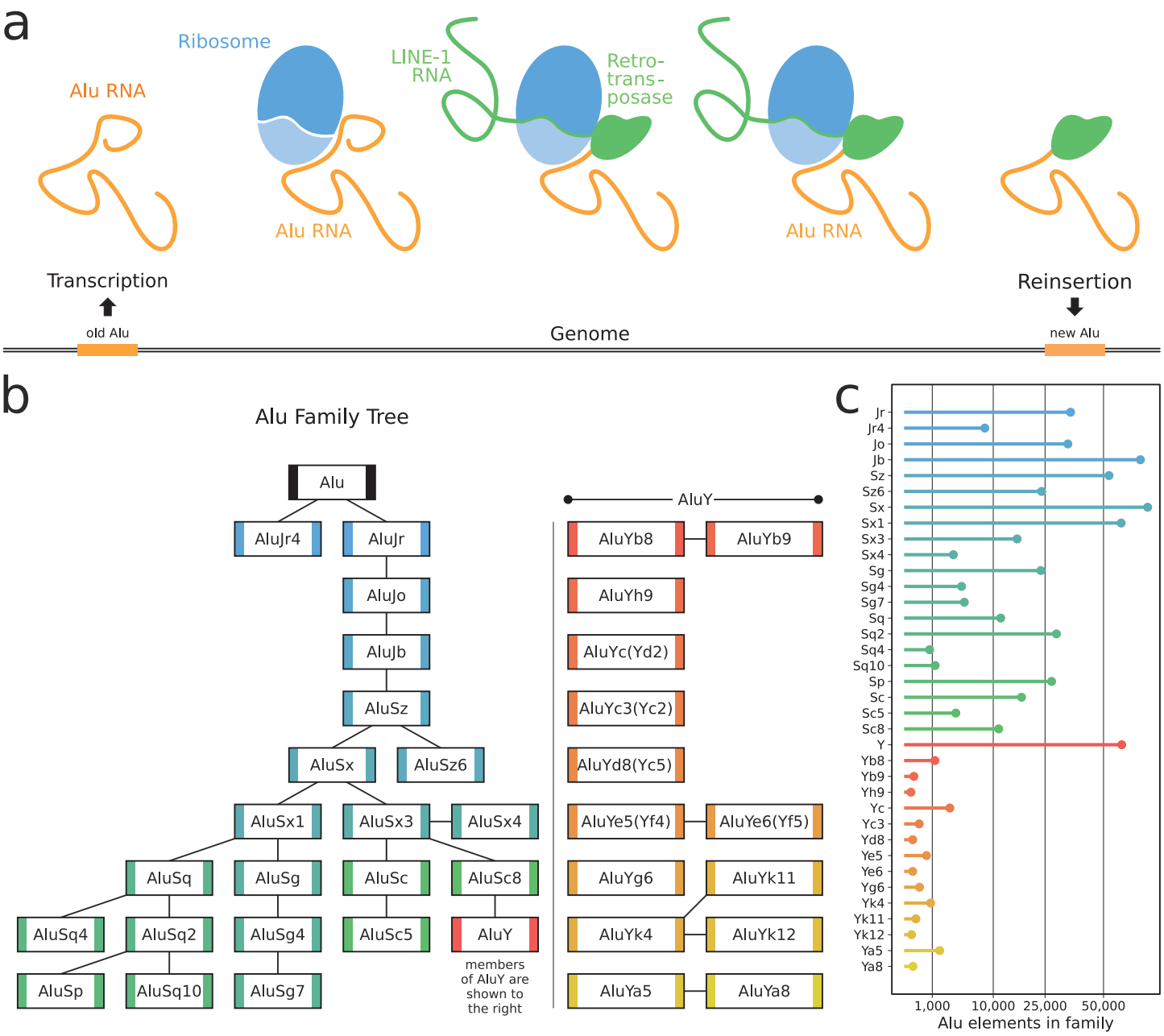

Figure 1 Alu evolution and retrotransposition **a**) Schematic representation of the Alu element reinsertion process (left to right). The Alu element is transcribed. – The Alu RNA attaches itself to the exit tunnel of the ribosome through its SRP sequence homolog. – A LINE-1 RNA arrives at the ribosome and its retrotransposase is translated. – The Alu RNA hijacks the LINE-1 retrotransposase. – The LINE-1 retrotransposase reinserts the Alu element into the genome at a new position. **b**) Human Alu family tree without designation of approximate evolutionary age (Price *et al.* 2004). Members of AluY are shown to the right. Alternative family naming conventions are denoted in brackets. The color scheme is reused in subsequent figures. **c**) Individual Alu Elements per family in the human genome (UCSC Genome Browser Annotation).

lated. It then uses the LINE-1 retrotransposase to reinsert itself into a new position in the genome (Häsler and Strub 2006).

The Alu life cycle carries with it certain risks for the host organism, as Alu element insertion into genes or other functional genomic regions can disrupt them (Deininger and Batzer 1999). As such, Alus have also been linked to increases in sequence diversity (Kazazian 2004; Ade *et al.* 2013), as well as influencing transcriptional activity in general (Chen and Yang 2017; Zhang *et al.* 2019), and under heat shock conditions in particular (Mariner *et al.* 2008). While generally lowly abundant, Alu elements are expressed and do successfully reinsert themselves into the human genome, with an estimated new Alu insertion for every 20 children born (Han *et al.* 2007; Xing *et al.* 2009; Conti *et al.* 2015).

While the exact evolutionary origin of Alu elements is unknown, they are assumed to be derived from the 7SL RNA, which is itself a component of the SRP (Ullu and Tschudi 1984; Kriegs *et al.* 2007). They possess a dimeric structure with a left and a right arm, separated by a variable A-rich region, as is also shown later in Figure 5 (Evgen'ev 2007). In addition, they possess two sequence features of note, the first being a bipartite RNA Polymerase III

(Pol-III) promoter located in the left arm, which is split into A box and B box (Paolella *et al.* 1983; Orioli *et al.* 2012). The second is the UGU(NR) binding motif required for attachment to the ribosome exit tunnel (Dagan *et al.* 2004).

While only present in higher primates, Alu elements most likely evolved from B1 repeats in rodents, which became the free left and right Alu monomers (FLAM and FRAM) in primates, which lastly fused to form the Alu elements (Quentin 1992; Kriegs *et al.* 2007). However, Alu evolution through error-prone retrotransposition continues and has given rise to many Alu families (Richard Shen *et al.* 1991; Deininger *et al.* 1992). The evolutionary oldest is AluJ, followed by AluS, and finally, AluY, which shows the most transposition activity in humans (Batzer *et al.* 1996). An Alu family tree is shown in Figure 1b (.json encoding in Supplement S1).

Despite the presence of a Pol-III promoter in their sequence, Alu elements are expected to not only be transcribed by Pol-III but also by RNA Polymerase II (Pol-II) (Conti *et al.* 2015; Zhang *et al.* 2019). In this study, we investigate different hypotheses regarding Alu transcript origins further, using metabolic RNA labeling by 4-thiouridine (dynamic transcriptome analysis, DTA, Schwalb *et al.* 2012) coupled with inhibition of Pol-II (see Methods, *Alu RNA* used interchangeably with *Alu transcripts*). This is of particular interest regarding past experiments that utilized Pol-II inhibition, as it is common practice to rely on SINEs as a non-Pol-II-dependent negative control group in such cases.

We also explore the half-life of Alu RNA in the cell, which is expected to be low, as the transcripts are presumed to be unstable (An *et al.* 2004). Finally, we also analyze the totality of annotated Alu loci to detect sequence features that influence Alu transcription. While standard motif search has been applied to this problem with limited success in previous studies (Zhang *et al.* 2019), we use a base-level generalized linear model (Nelder and Wedderburn 1972), and alternatively a colored and compacted de Bruijn graph (Idury and Waterman 1995; Pevzner *et al.* 2001; Iqbal *et al.* 2012), which lead to the *de novo* discovery of highly significant motifs that can partly be assigned to known transcription factors.

METHODS

Data Sets

The sequencing data used in this investigation is described in detail in our previous publication Schwalb *et al.* (2016), exception for the sequencing data relating to the inhibition of RNA Pol-II by α-amanitin, which was newly generated. Both data sets are available in NCBI's Gene Expression Omnibus, as described in Data Availability, including a table containing the positions, counts, and differential expression analysis results for the examined RNAs. The data set from Schwalb *et al.* (2016) is used throughout this investigation, except where α-amanitin inhibition is concerned.

K562 cells

Human K562 erythroleukemia cells were obtained from DSMZ (Cat. # ACC-10; RRID: CVCL_0004). K562 cells were cultured in accordance with the DSMZ Cell Culture standards in RPMI 1640 medium (Thermo Fisher Scientific) containing 10% heat inactivated fetal bovine serum (FBS) (Thermo Fisher Scientific), 1% Penicillin-Streptomycin and 1x GlutaMAX supplement (Thermo Fisher Scientific) at 37 °C in a humidified 5% CO_2 incubator. K562 cells used in this study display the phenotypic properties, including morphology and proliferation rate, that have been described in literature. Cells were verified to be free of mycoplasma contamination using Plasmo Test Mycoplasma Detection Kit (InvivoGen). Biological replicates were cultured independently.

α-Amanitin Treatment

α-Amanitin is a toxic substance from the mushroom *Amanita phalloides* and a potent inhibitor of RNA Pol-II (Lindell *et al.* 1970; Kedinger *et al.* 1970; Stirpe and Fiume 1967; Jacob *et al.* 1970). Among the RNA polymerases, RNA Pol-II is inhibited at low α-amanitin concentrations, RNA Pol-III may be inhibited at high concentrations ($> 250\,\mu g\,mL^{-1}$), while RNA Polymerase I (Pol-I) and mitochondrial RNA polymerases remain unaffected by α-amanitin. Treatment conditions were optimized for selective RNA Pol-II inhibition (see RT-qPCR and Western blotting). For α-amanitin (5 or $15\,\mu g\,mL^{-1}$) treatments of K562 cells, cells were treated for a time-course of 0 to 9 h at 37 °C in a humidified 5% CO_2 incubator. For TT-seq and RNA-seq, K562 cells were treated for 8 h at 37 °C in a humidified 5% CO_2 incubator with $5\,\mu g\,mL^{-1}$ α-amanitin or solvent (water).

RT-qPCR

For each condition (5 or $15\,\mu g\,mL^{-1}$ α-amanitin or solvent), 5×10^6 cells were harvested at 3000 x g for 2 min. Total RNA was isolated with QIAzol (QIAGEN) according to manufacturer's instructions except for the addition of 10 ng RNA spike-in mix (six spike-ins selected from ERCC RNA spike-in mix) (Schwalb *et al.* 2016) together with QIAzol. To remove possible genomic DNA contamination, isolated RNA (10 ng) was treated with TURBO DNase (Thermo Fisher Scientific) according to manufacturer's instructions. For reverse transcription (RT), random hexamer priming (5′- d(NNNNNN) - 3′, N = G, A, T, or C) was used according to manufacturer's instructions. Briefly, 1 µg of DNase-treated RNA, Random Hexamer primers (final concentration of $5\,ng\,\mu L^{-1}$), dNTPs mix (final concentration of 0.5 mM) were mixed and incubated at 65 °C for 5 min. Subsequently, Maxima H Minus Reverse Transcriptase (RT) (final concentration of 200 U) and 5x Maxima RT buffer (Thermo Fisher Scientific) were added (+RT reaction). For DNA contamination control, cDNA synthesis without RT (−RT reaction) was performed (RT was substituted with water). The (−/+) RT reactions were incubated in a PCR cycler at 25 °C for 10 min, 50 °C for 30 min, and 85 °C for 5 min. Primers for quantitative PCR (qPCR) were designed by using the online primer design software Primer3 v.0.4.0 (Rozen and Skaletsky 2000). Briefly the selection of Pol-II targets was based on half-lives measurements in K562 cells (Schwalb *et al.* 2016). Primers were targeted to Pol-II transcripts (PAIP1, CWC22, EGR1) as a positive control, RNA Pol-I transcripts (18S rRNA) as a negative control, and RNA Pol-III transcripts (U6 snRNA). Primer specificity (single product peak) was validated by melting profiles. Primer sequences, length, PCR efficiency values of primers (E) and targets are reported in Supplement S2. cDNAs (50 ng) were amplified with SYBR Select Master Mix (Thermo Fisher Scientific) according to manufacturer's instruction with a final primer concentration of 400 nM. PCR reactions were run in 96-well optical plates sealed with optical adhesive cover on a qTOWER 2.0 / 2.2 instrument (Analytik Jena AG). The following thermal cycling conditions were used (SYBR Select Master Mix reference, standard cycling mode): 50 °C for 2 min, 95 °C for 2 min, 40 cycles of 95 °C for 15 s, and 60 °C for 1 min. Two synthetic RNA spike-ins were used for normalization. The $2^{-\Delta\Delta Ct}$ method was applied to calculate the normalized target gene expression fold change, with the amplification efficiency (E) for each target gene,

slope of standard curve (S) and mean threshold cycle (Ct) (Livak and Schmittgen 2001).

Western Blot

For α-amanitin (5 or 15 µg mL^{-1}) treatments of K562 cells, cells were treated for 8 h at 37 °C in a humidified 5% CO_2 incubator. Water was used as a solvent. After 8 h treatment, K562 cells were harvested at 3000 x g for 2 min, and washed twice in DPBS. Cells were lysed in RIPA lysis buffer for 45 min on ice and centrifuged at 14 000 x g for 15 min at 4 °C. Samples were quantified with the Bradford method. 15 µg cell lysate was denatured in 4X Loading Dye (including 100 mM DTT) at 70 °C for 10 min. PrecisionPlus Protein All Blue Standard (Bio-Rad, #161-0373, 10-250 kDa) was used as marker. NuPAGE 4-12% Bis-Tris Protein Gels and MOPS buffer were used according to manufacturer's instructions. Transfer of a single NuPAGE onto PVDF membrane was performed in transfer buffer at 30 V for 1 h in a XCell II Blot Module (semi-wet transfer unit) according to manufacturer's instructions. Membrane blocking was performed in 5% milk PBS-T on a rocking surface for at least 1 h. Primary antibody was added overnight. The following primary antibodies were used in this study: N-terminal POLR2A / hRPB1 antibody, clone F-12 (Santa Cruz Biotechnology, sc-55492, Lot. # E2913; RRID: AB_630203) and GAPDH antibody, clone C 71.1 (Sigma-Aldrich, G8795, Lot. # 067M4785V; RRID: AB_1078991) as a loading control. HRP-coupled secondary antibody targeting mouse IgG (Abcam, ab5870) was used at a dilution of 1:3000 and incubated on a rocking surface for 1 h. ECL working solution of SuperSignal West Pico PLUS Chemiluminescent Substrate was prepared according to manufacturer's instructions. Proteins were visualized by chemiluminescence detection on INTAS. POLR2A / hRPB1 (RNA Pol-II) degradation which is α-amanitin dose dependent (Nguyen *et al.* 1996) was monitored by Western blotting.

RNA Spike-Ins

Synthetic RNA spike-in controls are derived from selected RNAs of the ERCC RNA Spike-in Mix (Ambion) as described in Gressel *et al.* (2019b) and Schwalb *et al.* (2016). Briefly, spike-ins (three unlabeled and three 4sU-labeled) are *in vitro* transcribed using the MEGAscript T7 kit (Ambion). *In vitro* transcription (IVT) of unlabeled spike-ins was performed following manufacturer's instruction. For IVT of 4sU-labeled spike-ins, 10% of UTP was substituted with 4-thio-UTP (Jena Bioscience). RNA spike-ins were purified with RNAClean XP beads (Beckman Coulter) following manufacturer's instructions. The final RNA spike-in pool contained equal amounts of all RNA spike-ins.

TT-seq and RNA-seq

A detailed step-by-step protocol has been deposited in the protocols.io repository (Gressel *et al.* 2019a). TT-seq and RNA-seq were performed in two biological replicates including RNA spike-ins. Briefly, experiments were performed using 5×10^7 K562 cells per biological replicate. Cells were kept at optimal growth conditions and supplemented with 5 µg mL^{-1} of α-amanitin or solvent (water) for 8 h. After 7 h 55 min, a 4-thiouridine (4sU) labeling pulse (Sigma-Aldrich, T4509) was applied for 5 min using 500 µM (see Supplement S3). Total RNA was isolated with QIAzol reagent (QIAGEN, #79306) according to manufacturer's instructions except for the addition of 150 ng RNA spike-in pool with QIAzol reagent as previously described (Gressel *et al.* 2019a; Schwalb *et al.* 2016). The Ovation Universal RNA-Seq System (NuGEN) was used for strand-specific library preparation as described (Gressel *et al.* 2019b). Purified cDNA libraries were analyzed by Fragment Analyzer prior to Illumina sequencing. Sequencing was performed on a HiSeq 2500 (Illumina) in paired-end mode with 50 bp read length.

TT-seq / RNA-seq Prepossessing and Normalization Parameters

TT-seq and RNA-seq prepossessing and normalization were performed as detailed in Schwalb *et al.* (2016), with an alternative normalization applied as described under Differential Expression Analysis. Reads were mapped with STAR v2.7.3a (Dobin *et al.* 2016), allowing a maximum of 10 mismatches and minimal uniqueness filtering with a MAPQ value cutoff of 255. The general error rate per library ranged between 0.6 % and 0.9 %. As Alu sequences are notoriously difficult to map (Sexton and Han 2019), we performed a mappability analysis (see Supplement S4) to make certain that our read alignment does not introduce biases. We simulated a homogenious coverage of the genome and mapped the simulated reads using the same aligner settings. Our analysis shows that there is no bias caused by Alu mappability, which is also corroborated by the findings of Sexton and Han (2019), who show that the mappability of transposable elements can be improved through the use of paired-end read libraries to the point where a majority of elements are uniquely mappable. Association of the sequencing data with specific Alu elements or protein-coding genes used the GRCh37 / hg19 repeat annotation or respectively the canonical gene annotation of the UCSC Table Browser (Karolchik *et al.* 2004), in addition to the RepeatMasker human Alu subfamily re-analysis (Price *et al.* 2004). We also tested using the transcription unit (TU) annotation from Gressel *et al.* (2019b) based on the GenoSTAN segmentation algorithm as an alternative to the UCSC canonical gene annotation, which leads to a similar mRNA read count distribution (Zacher *et al.* 2017, data not shown).

RNA Half-Life Estimation

To estimate the half-life of expressed Alu elements and coding transcripts we used MLE, described in detail in Supplement S5. Our estimation makes use of the labeled and total RNA sequencing fractions obtained through TT-seq and 4sU-seq, as detailed above. TT-seq and 4sU-seq data were both used for the estimation, as the correlation between the two methods is suitably strong ($r > 0.80$ between replicates and $r > 0.85$ between sequencing methods). While 4sU-seq was designed to estimate half-lives, TT-seq was primarily designed to measure polymerase processivity during transcription. Since Alu elements are short, TT-seq essentially measures the synthesis rate of Alu transcripts and can thus be applied for half-life estimation (see Supplement S6).

We assume that the total amount of transcripts in a cell remains constant and that transcript degradation follows exponential decay, meaning that the ratio of labeled to total transcripts L/T increases exponentially to 1 over time. We further assume that the distribution of labeled and total counts follow a Poisson distribution. This allows us to formulate the likelihood function $\mathcal{L}$ of observing a specific number of labeled counts L_a and total counts T_a for any given Alu element or coding transcript a, as a parametrized combination of the underlying Poisson distribution:

$$\mathcal{L}(L_a, T_a; r_a, q, L_{spk}, T_{spk}) = \text{Pois}(T_a; \lambda_{tot} = c_{tot} \cdot r_a^{-1}) \cdot \text{Pois}(L_a; \lambda_{lab} = c_{lab} \cdot r_a^{-1})$$

Here r_a is the ratio between labeled and total molecules for any given Alu element or coding transcript a; q is the ratio between labeled and total spike-in molecules that were added; L_{spk} and T_{spk}

are the number of labeled and total spike-in counts respectively; and c_{lab} and c_{tot} are given as:

$$c_{tot} := \frac{L_a}{L_{spk}} \cdot T_{spk} \cdot q \quad \text{and} \quad c_{lab} := L_{spk} \cdot \frac{T_a}{T_{spk}} \cdot \frac{1}{q}$$

This estimation does not depend on feature length, as we use the ratio between labeled and total counts. To find the optimal estimate of r_a, we maximize the log likelihood function:

$$\ell(r_a; L_a, T_a, q, L_{spk}, T_{spk}) = L_a \left(\ln(r_a) - \frac{T_{spk}}{L_{spk}} \frac{q}{r_a} \right) + T_a \left(-\frac{L_{spk}}{T_{spk}} \frac{r_a}{q} - \ln(r_a) \right)$$

From r_a when then obtain the degradation rate δ_a by use of $r_a = 1 - \exp(-\delta_a \Delta t)$, where Δt is the labeling pulse's duration. This allows us to calculate the half-life as $t_{1/2} = \ln(2) / \delta_a$

As the used DTA methods rely on the labeling of U as 4sU, the percentage of U in an Alu sequence could potentially bias its half-life estimate. However, we can exclude this possibility, as no correlation exists between Alu half-life and U ratio ($r < 0.01$).

Differential Expression Analysis

Differential expression analysis of the α-amanitin Pol-II inhibition experiment was performed using the DESeq2 package for R (Love *et al.* 2014), using un-normalized counts as required by the statistical model. The standard normalization strategy employed by DESeq2 internally relies on the assumption that there are no substantial, systematic global changes in (mRNA) expression between samples. Since this is most likely not given in the α-amanitin sample, we used a set of *bona fide* housekeeping transcripts. As it is known that the activity of mitochondrial polymerase is unaffected by α-amanitin (Menon 1971; Reid and Parsons 1971; Saccone *et al.* 1971), we chose mtRNAs for normalization. Alternatively, to check the robustness of our choice, we also performed normalization based on a set of spike-in RNAs that had been added during sequencing library preparation (see RNA Spike-Ins). Similar results were obtained when spike-in RNAs were taken as reference as compared to mtRNAs (Supplement S7c). The normalization constants obtained differ by a mean factor of 1.10, and consequently the results of our analysis hold under both regiments.

Rejection sampling was used to ensure that the obtained l2fc distributions are comparable (Flury 1990). To avoid biases arising from the substantially different expression level distribution of Alus and mRNAs, we compare the Alu l2fc distribution to the l2fc distribution of a large sample of mRNAs that were picked randomly such that their wild type expression distribution is identical to that of the Alu elements. To generate this random sample, we use rejection sampling: We define expression breakpoints

$$b_0 = -\text{inf} < b_1 < b_2 < \ldots < b_{N-1} < b_N = \text{inf}$$

such that each bin $[b_n, b_{n+1}]$ contains 1500 Alu elements, excluding Alu elements with less then 5 reads. Thus the empirical expression distribution

$$f(n) = p\left(\text{Alu expression in } [b_{n-1}, b_n]\right) = 1/N, \quad n = 1, \ldots, N$$

is uniform. The same breakpoints are used to calculate the empirical expression distribution

$$g(n) = p\left(\text{mRNA expression in } [b_{n-1}, b_n]\right), \quad n = 1, \ldots, N$$

of the mRNAs by calculating the relative abundance of mRNAs whose expression falls into the interval $[b_{n-1}, b_n]$. Let $k = \min\left(g(n), n = 1 \ldots, N\right)$ and accept an mRNA sample that lies in the interval m with probability $k/g(k)$. It can then be shown that the set of accepted samples approaches, for large numbers, the distribution $f(n)$.

Generalized Linear Model

The GLMs to analyze the genomic sequences of all annotated Alu elements on base-level were created with the glmnet package for R (Friedman *et al.* 2010). A Poisson family distribution response type was assumed with observation standardization, an elastic mixing parameter α of 1 (full Lasso penalty, no ridge regression penalty), no fitted intercept parameter, and 1000-fold cross-validation. The input matrix for each model consisted of a binary encoding of the examined point mutation, base exchange, deletion or insertion, for each position in the Alu consensus sequence. These matrices were obtained by aligning each individual Alu sequence against the Alu consensus sequence. These predictor matrices where then paired with the expression values as response variable for each individual model. The resulting effect sizes (β parameters) for the point mutations were then annotated along the Alu secondary structure according to Sinnett *et al.* (1991) (dot-bracket notation of the secondary structure in Supplement S8) to illustrate their context regarding Alu sequence features. The β parameters are the estimated regression coefficients of the GLM.

De Bruijn Graph

The de Bruijn graph of all Alu sequences, regardless of expression, was generated using bifrost v1.0.5 (Holley and Melsted 2020), with a k-mer length of 9 and including 100 bp of flanking regions up- and downstream of each Alu sequence. k = 9 was chosen so that the distribution of Alu sequences per node in the graph was smooth, neither running into the lower nor the upper range of possible values (Supplement S9). Low coverage k-mer connecting tips were kept by using the `--keep-mercy (-y)` argument, and the compacted de Bruijn graph was colored using the `--colors (-c)` argument.

Our initial intent was to utilize the de Bruijn graph structure in the downstream analysis, but the compacted de Bruijn graph for k = 9 is almost complete: It contains 131 070 nodes, i.e., k-mers and 1 048 544 in-edges, resulting in > 7.99 edges per node (similarly > 7.99 for out-edges). Note that the maximum in- and out-degree in the compacted graph is eight and not only four since each node can be traversed in forward and reverse complement direction (Holley and Melsted 2020). We, therefore, focused purely on the k-mers.

The resulting k-mers were then filtered according to two aspects. Firstly, the expression of those Alu elements possessing the k-mer was compared with those not possessing it. Secondly, for any given k-mer, its suffixes and prefixes were generated. Each k-mer possesses 4 potential pre- and suffixes, as each sequence can be preceded or succeeded by one of the four bases.

When testing a k-mer, say XJY, with $X, Y \in \{A, C, T, G\}$ and J a fixed 2-mer, for having an effect on Alu transcription, we have to guard against false positive due to the large number of k-mer tests. This is done by Bonferroni multiple testing correction for a family-wise error rate of at most 5%. We further need to filter out irrelevant findings resulting from the large number of observations that can render even small differences highly significant. To ensure that XJY is relevant for Alu transcription, we compare the group of Alu elements containing XJY to the group of Alu elements

that contain the prefix XJ or the suffix JY, but not the full k-mer XJY. We compare the two groups with respect to their binarized expression (expressed with at least one read count / not expressed) using a Fisher test, and we require an OR of at least 2 or at most 0.5 for being considered a relevant difference.

RESULTS

Alu expression increases with family age

The many Alu families that have emerged in the human genome over the last 65 million years vary not only in their sequence characteristics (Jurka and Smith 1988), but also in their activity (Bennett *et al.* 2008; Oler *et al.* 2012); active Alu elements being those that are not only transcribed but are also mobile, i.e., have the capability to reinsert themselves into the genome through retrotransposition.

As was found by Bennett *et al.* (2008) *in vitro*, the activity of an Alu family appears to scale inversely with its age; the older an Alu element, the less likely it is to be still capable of successful reinsertion. We examined the transcriptional activity of Alu families, using the RNA-seq data from Schwalb *et al.* (2016) (see Methods), as well as the stability of Alu transcripts, which is expected to be low in comparison to regular mRNA (An *et al.* 2004 and Figure 2a-e). We used maximum likelihood estimation (MLE) based on metabolic RNA labeling to calculate the half-life $t_{1/2}$ of Alu RNAs (see Methods), which relates to the degradation rate δ by $t_{1/2} = \ln(2)/\delta$, as further detailed in Supplement S5.

The global expression distribution of Alu transcripts (mean $\bar{x} \approx 85$, median $\tilde{x} = 27$) is about two orders of magnitude lower than that of mRNAs ($\bar{x} \approx 4088$, $\tilde{x} = 740$), as is to be expected (Figure 2a, Paulson and Schmid 1986). We found that members of the younger Alu families, mostly AluY, but also AluSq10, exhibit overall less transcriptional activity than older Alus (Figure 2b). This effect is most pronounced with the subfamilies of AluY. Since the differences between the heavy-tailed read count distributions are not easy to visualize, we compare the empirical cumulative distribution functions (ECDF) of the young Alu families with the ECDF of the old ones (Figure 2e). All young ones are significantly different from the old ones ($p < 0.05$ in all comparisons, Kolmogorov-Smirnov (KS) test or Wilcoxon test, both with Bonferroni multiple testing correction). The split can be further verified by performing average linkage hierarchical clustering with the KS test statistic as a distance measure between Alu families. This separates the AluY subfamilies clearly from the older families, except for the aforementioned AluSq10 and AluYc3 (AluYc2) (Supplement S7a).

The use of TT-seq (Schwalb *et al.* 2016) and 4sU-seq (Miller *et al.* 2011) allows for the identification of newly created transcripts, and thereby the estimation of RNA synthesis and degradation rates, and finally half-lives (see Methods). However, due to the overall paucity of Alu read counts, our MLE results are relative and do not represent explicit values (see Discussion). Our half-life estimates for Alu elements show a dispersion very similar to that of mRNAs (Figure 2c). This result is surprising, as the *in silico* study by An *et al.* (2004) suggests that Alu transcripts should be notably less stable

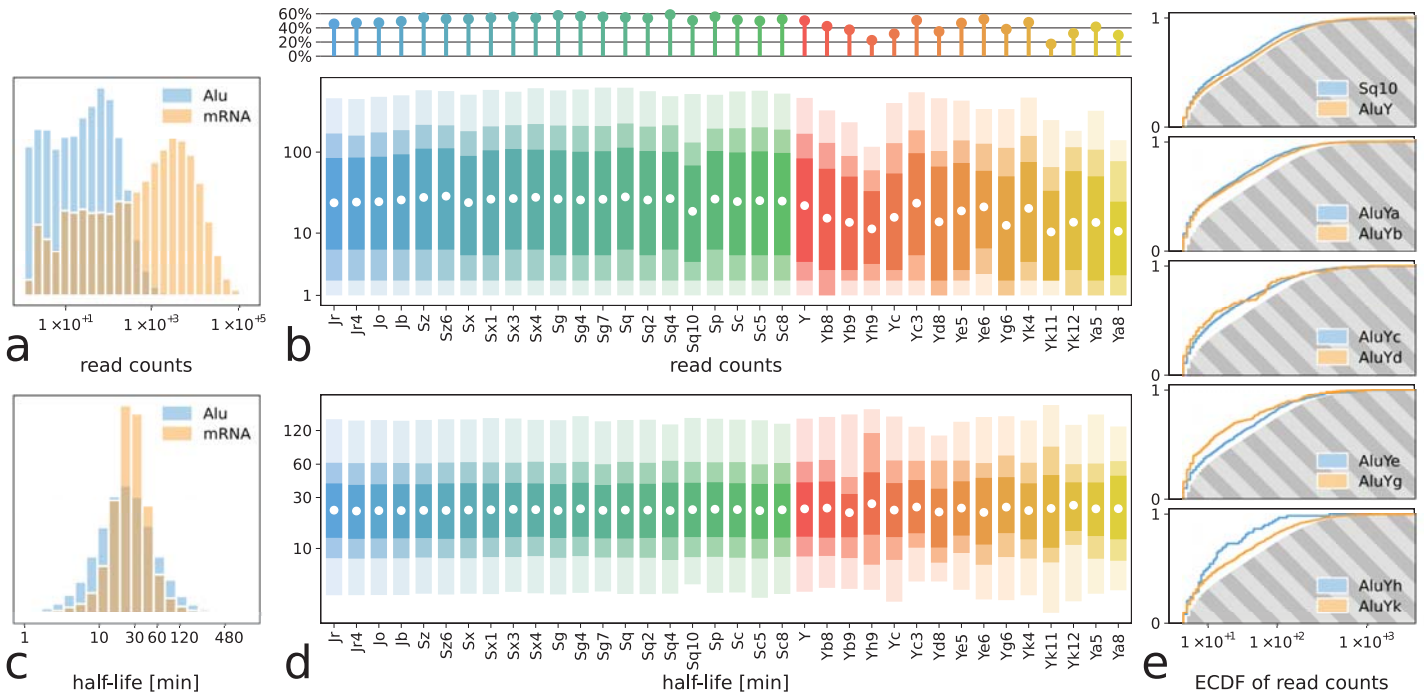

Figure 2 Alu expression and half-life – All subfigures based on data published by Schwalb *et al.* (2016), using the UCSC Genome Browser Annotation. **a**) Histogram of the read counts for both Alu elements and mRNA; Alu elements show less transcriptional activity overall. (histogram y-axes: density) **b**) Read count distribution for all Alu families as shown in Figure 1b. The colored bars represent the extent of the 50%, 75%, and 95% median percentile interval with decreasing opacity. The white dot indicates the arithmetic mean. Zero counts are disregarded in this subfigure. The upper lollipop plot shows the proportion of Alu elements per family with more than 0 read counts. **c**) Histogram of the half-life of both Alu elements and mRNA; Alu elements show stability comparable to regular transcripts. **d**) Half-life distribution for all Alu families (cf. Figure 1b), binned by family. **e**) ECDF of the read counts of younger Alu families (AluY and subsequent, as well as AluSq10) in contrast to all older families, represented by the gray shaded area. Younger Alu families show consistently lower read counts in comparison to the older ones.).

than mRNAs. Additionally, the half-life distribution of Alu RNAs remains similar between families, even in the youngest subfamilies of AluY that have only a few members and are generally lowly expressed (Figure 2d).

Alu expression is independent of gene transcription

Alu elements are often presumed to be Pol-III transcripts, as they contain a bipartite Pol-III promoter (Orioli *et al.* 2012). Although there is experimental evidence for Alu elements being transcribed by Pol-III (Zhang *et al.* 2019; Panning and Smiley 1993; Jagadeeswaran *et al.* 1981), it has not been ruled out yet that (some) Alu elements might be transcribed by RNA Polymerase II (Pol-II). We have therefore examined the possibility that Alu transcripts may also arise as side products of RNA Pol-II gene transcription, as hinted at by the results of Conti *et al.* (2015) and suggested by Zhang *et al.* (2019).

If Alu elements are also transcribed by Pol-II as side products of gene transcription, we would expect intragenic Alu elements to exhibit a higher transcript abundance than intergenic ones. Further, the transcriptional activity of genes should be correlated with that of proximal Alu elements. The expression of those should be

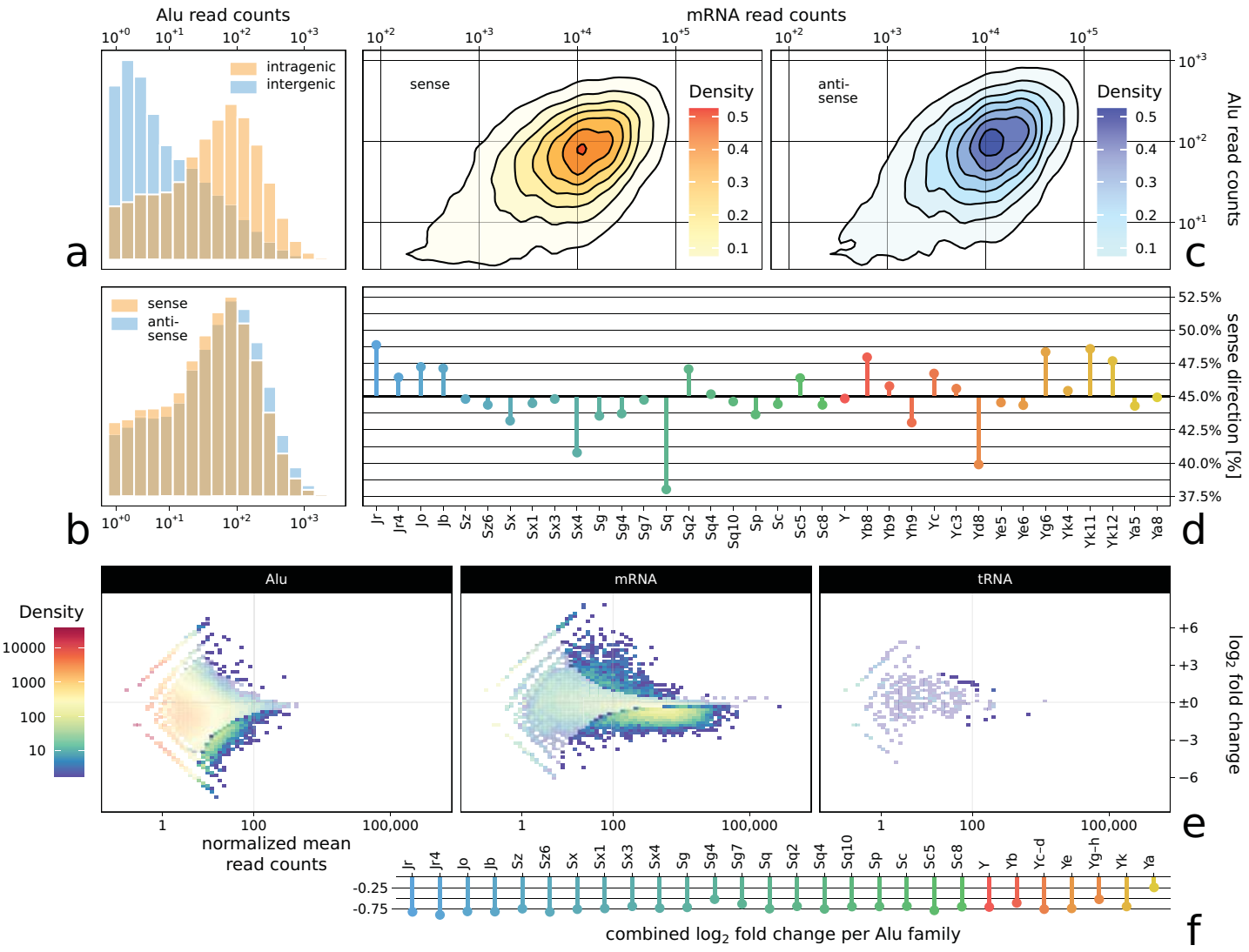

Figure 3 Alu correlation and differential expression **a**) Histogram of the read counts for transcribed intragenic (398 660) and intergenic (200 928) Alu elements. Intragenic Alu elements either directly overlap a gene or lie within 500 bp up- or downstream of a gene. Other Alu elements are intergenic. Intragenic Alu elements show a significantly higher expression than intergenic ones, which is likely explained by genome accessibility. **b**) Histogram of the read counts for transcribed intragenic Alu elements that lie in sense or antisense direction with respect to their associated gene, showing only minor differences. **c**) 2D kernel density estimation showing the correlation between the annotated Alu and mRNA read counts for intragenic sense (orange) and antisense (blue) Alu elements. The two groups show no difference in correlation strength. **d**) Number of sense and antisense Alu elements per family. No family exceeds a ratio of 38% or 50%. **e**) 2D density heatmap showing the DESeq2 differential expression of Alu elements, mRNAs, and tRNAs as control under α-amanitin Pol-II inhibition. Semi-transparent areas do not pass the significance threshold. Loci with a normalized mean expression below 0.1 are not shown, with affects 51% of all annotated Alu loci and practically no mRNAs or tRNAs. Both Alu elements and mRNAs show stronger significant down- than upregulation. **f**) Differential expression of Alu elements under α–amanitin Pol-II inhibition, using family-wise aggregated read counts. All Alu families show downregulation.

biased towards Alu elements lying in sense direction with regard to their associated gene.

To investigate this, we defined intragenic Alu elements as those that either directly overlap a gene or lie within 500 bp up- or downstream of a gene according to the GRCh37 / hg19 canonical gene annotation of the UCSC Table Browser (Karolchik *et al.* 2004). All other Alu elements are called intergenic. Over 96 % of intragenic Alu elements overlap intronic regions according to the latest GENCODE annotation (Frankish *et al.* 2021). Effects of proximity to Pol-I and Pol-III transcripts have not been detected and cannot contribute a notable effect, since only a comparably tiny number of Alu elements would be affected. Of the expressed Alu elements, 37 loci lie in 500 bp proximity to rRNAs (excluding 5S rRNA), and 42 in proximity to tRNAs, using the respective UCSC Table Browser annotation.

We compared the read counts for intragenic and intergenic Alu transcripts, which show a significantly higher expression for intragenic Alus compared to intergenic ones (Figure 3a, $\bar{x} \approx 110$, $\tilde{x} = 49$ intragenic, $\bar{x} \approx 35$, $\tilde{x} = 6$ intergenic, $p < 10^{-5}$, Mann-Whitney test). We also detected a less pronounced but significant difference in expression of intragenic Alu elements lying in sense respectively in antisense direction with regard to their associated gene (Figure 3b, $\bar{x} \approx 97$, $\tilde{x} = 44$ sense, $\bar{x} \approx 121$, $\tilde{x} = 54$ antisense, $p < 10^{-5}$, Mann-Whitney test)

We also correlated the transcript abundance of intragenic Alu elements and their associated genes for both Alu elements in sense and antisense direction to their associated gene. The two groups show no difference in correlation strength (Spearman correlation $r \approx 0.71$, $p < 10^{-5}$ rand. test, Figure 3c).

Lastly, we examined the number of sense and antisense Alu elements per family. Overall, 46% of all intragenic Alu elements lie in sense direction with regard to their associated gene. No individual family exceeds a ratio of either 38% or 50% (Figure 3d).

Furthermore, we combined Pol-II and Pol-III occupancy data drawn from the ENCODE portal (ENCSR000EHL, ENCSR000EHQ, Dunham *et al.* 2012; Davis *et al.* 2018) with our RNA-Seq data and RNA-Seq data from ENCODE (ENCSR000COM). While we found a modest correlation between our own sequencing's mRNA transcript abundance and the data obtained from ENCODE (Spearman correlation $r \approx 0.65$), our analysis did not detect global correlation between Pol-II or Pol-III occupancy and individual Alu expression according to our sequencing data (Spearman correlation $r \leq 0.05$ for both, Supplement S7b).

RNA Pol-II inhibition decreases Alu expression

To further examine the origin of Alu transcripts, we performed a Pol-II inhibition experiment by incubating K562 cells with α-amanitin at a concentration suited to efficiently inhibit Pol-II activity while leaving the activity of Pol-III and mitochondrial RNA polymerases unaffected (see Methods). We quantified the degree of inhibition using mitochondrial RNAs as a negative control group which are unaffected by a-amanitin (see Methods). Similar results were obtained when spike-in RNAs were taken as reference (Supplement S7c). The analysis of tRNAs as an independent negative control group showed barely any differential expression, validating that our normalization procedure did not introduce a bias to putatively unaffected transcripts. As a natural positive control group, we chose mRNAs that are *bona fide* Pol-II transcripts. Figure 3e shows the changes after α-amanitin inhibition for Alus elements, mRNAs, and tRNAs.

Due to the low coverage of individual Alu transcripts, a direct comparison based on individual expression folds would be misleading. We, therefore, aggregated all read counts mapping to one Alu family (or several families in the case of very small families, Figure 3f) and calculated the expression fold of each Alu family. All Alu families show downregulation with an average $\log_2$ fold change (l2fc) of -0.70 (standard deviation $\sigma \approx 0.12$, min $= -0.88$ for AluYr4, max $= -0.24$ for AluYa). In comparison, individual mRNAs with at least 10 reads exhibit a downregulation of -0.35 ($\sigma \approx 0.99$). Also, Alu elements being intragenic or intergenic shows no influence on their response to α-amanitin (>90% downregulated among significantly differentially expressed transcripts in both cases), therefore these two groups are treated as one in the following.

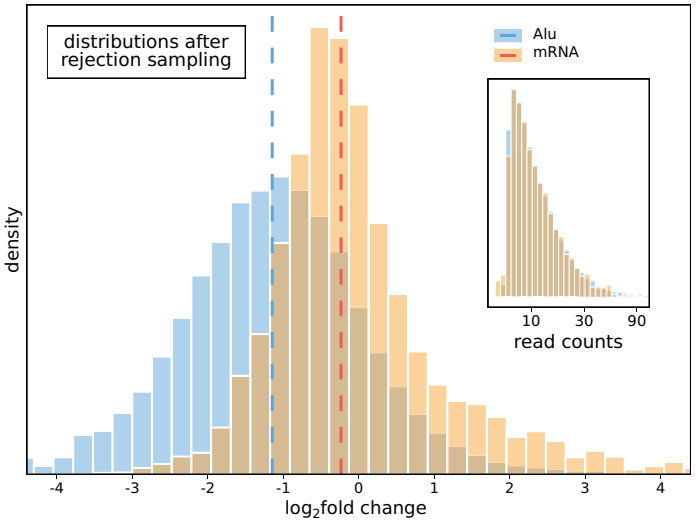

Figure 4 Rejection sampling – Distribution of DESeq2 differential expression of Alu elements and mRNAs after rejection sampling according to transcript length corrected Alu normalized mean read counts (shown in inset). Vertical lines denote median Alu l2fc of -1.15 (blue), and median mRNA l2fc of -0.24 (red). Alu elements exhibit downregulation comparable to canonical genes under α-amanitin Pol-II inhibition.

As the overall mean expression of Alu elements differs strongly from that of mRNAs (cf. Figure 2a), we used rejection sampling to obtain comparable l2fc distributions (Figure 4, Flury 1990). Alu and mRNA differential expression appears very similar, with the distribution of Alu l2fc falling even below that of mRNA. Examining individual l2fc values, we focused on those Alu elements or mRNAs with significant differential expression ($p \leq 0.05$, Wald test with Benjamini-Hochberg multiple testing correction). Of all Alu elements, 16 909 (89%) of 19 027 transcripts show no significant differential expression while 2118 (11%) do. 9818 (53%) of 18 590 mRNAs show no significant differential expression while 8772 (47%) do. tRNAs are largely unaffected by α-amanitin, with only 17 of the total 604 (3%) tRNA transcripts exhibiting significant change, 12 being up- and 5 being downregulated. This is to be expected, as tRNAs are transcribed by Pol-III (White 1997).

Alu elements, while expectedly represented with considerably less read counts than mRNA transcripts, still show downregulation under α-amanitin treatment, with an average l2fc downregulation of -3.23 ($\sigma \approx 1.24$) for significant transcripts. Respectively, mRNAs exhibit an average significant downregulation of -0.99 ($\sigma \approx 0.43$).

However, both Alu elements and mRNAs show a small fraction of significantly upregulated transcripts under α-amanitin in-

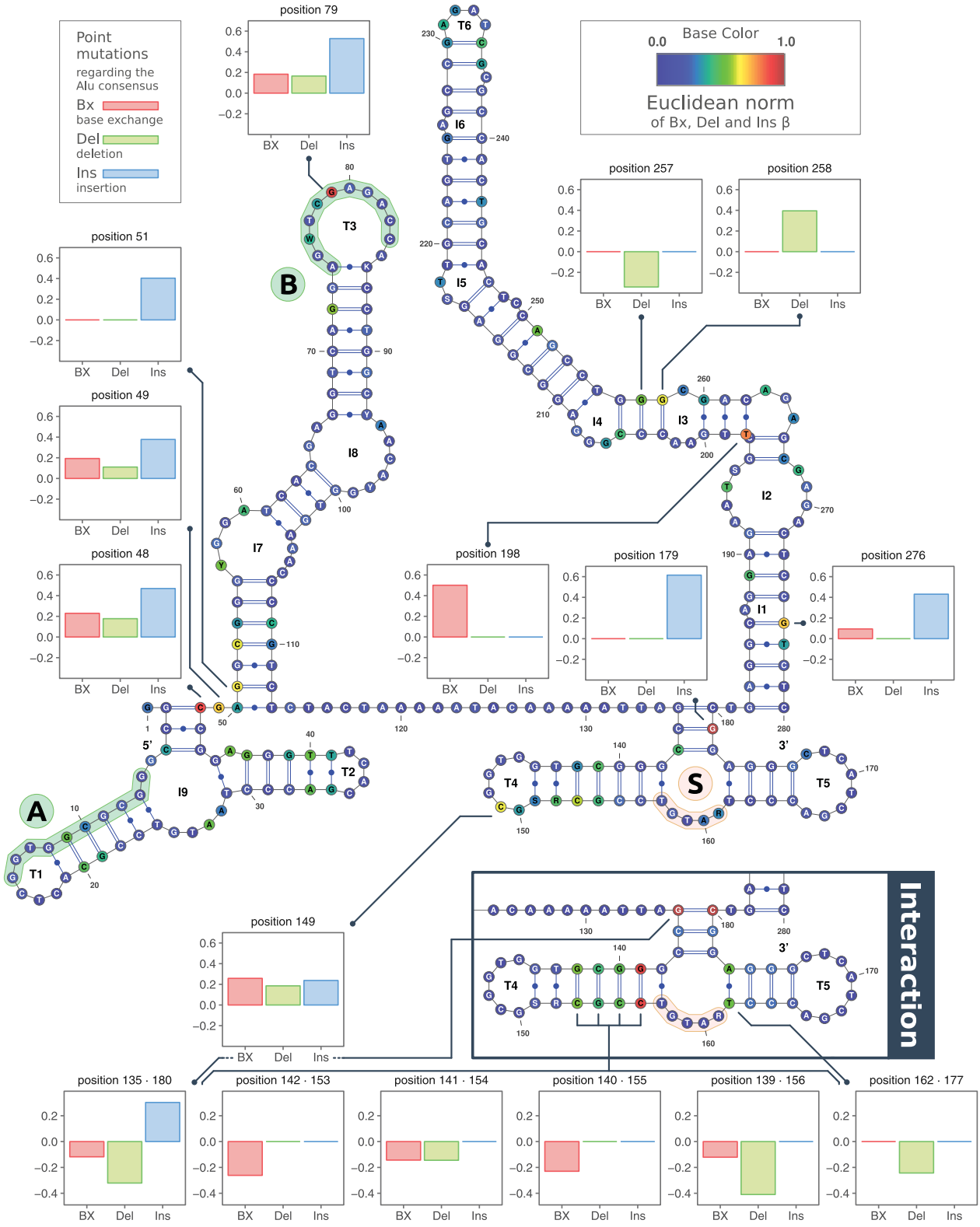

Figure 5 GLM and Alu secondary structure – Secondary structure of the Alu element consensus sequence according to Sinnett *et al.* (1991), showing the left and right arm architecture with the variable region in between. Noteworthy sequence features include the bipartite Pol-III promoter split into A and B box (denoted by Ⓐ and Ⓑ), and the SRP binding motif used for attachment to the ribosome exit tunnel (denoted by Ⓢ). Each base is colored according to the rescaled euclidean norm of the three GLM effect sizes (β parameters). Several noteworthy positions are annotated with bar graphs detailing the individual values. In the lower right corner, prominent results from the analysis of interactions between paired bases are shown.

hibition of Pol-II. In comparison to the significantly downregulated fraction though, this fraction is small with 1002 (5%) significantly upregulated but 7770 (42%) significantly downregulated transcripts for mRNAs, and 47 (0.2%) significantly upregulated but 2071 (11%) significantly downregulated transcripts for Alu elements.

It is remarkable that some Alu elements show significant downregulation under α-amanitin treatment at all, which suggests that the affected Alu elements could be partly Pol-II transcripts.

Sequence features associated with Alu expression show signs of evolutionary selection

Evolutionary changes in the Alu sequence could reveal informative clues on sequence features that influence the mechanism of Alu transcription and retrotransposition. Obvious, prominent features of potential impact are the Pol-III promoter split into A and B box and the SRP binding motif used for attachment to the ribosome exit tunnel (see Introduction).

The abundance of Alu sequences in the genome offers the unique opportunity to use an analysis strategy resembling a genome-wide association study (GWAS). We used a generalized linear model (GLM) to link Alu sequence features to their transcript abundance (see Methods). Due to the encoding of the sequence changes with respect to the Alu consensus sequence, this results in individual effect sizes (β parameters) for the three types of examined point mutations, base exchanges, deletions, and insertions, for each position in the Alu consensus sequence. Figure 5 shows the Alu secondary structure according to Sinnett *et al.* (1991) with the euclidean norm of the effect sizes represented by base color (dot-bracket notation of the secondary structure in Supplement S8). However, the large number of available Alu sequences enabling this analysis also precludes any attempts to test for statistical significance.

Several noteworthy positions show a high correspondence to Alu transcription (Figure 5): Position 48 is sensitive to all three types of point mutations, but primarily to insertions. It lies in the left arm at the juncture between inner loop 7 (I7) and inner loop 9 (I9), which contains the A box of the Pol-III promoter sequence. Position 79, also in the left arm in terminal loop 3 (T3), lies directly within the B box of the Pol-III promoter sequence. Like the previous position, it is sensitive to all three types of mutation, but primarily to insertions, which would bring the Alu consensus sequence closer to the consensus sequence of the tRNA promoter: GWTCRANNC (Paolella *et al.* 1983). Position 179 lies in the right arm near terminal loops 4 and 5 (T4, T5) and reacts only to insertions, and position 198, also in the right arm between inner loops 2 and 3 (I2, I3), reacts only to base exchanges.

Due to the large number of samples (i.e., Alu elements), we could also investigate interactions between paired bases in the secondary structure (see Methods), which resulted in two positions of interest, all in the vicinity of terminal loops 4 and 5 (T4, T5) of the right arm, which is also the region containing the SRP 9 / 14 binding motif: UGU(NR) (Weichenrieder *et al.* 2000). The pair of positions 135 and 180 reacts inversely to insertions in comparison to base exchanges and deletions. It forms the junction from which terminal loops 4 and 5 (T4, T5) branch off. The pair of positions 139 and 156 lies directly adjacent to the SRP binding motif and reacts primarily to deletions, as well as to base exchanges.

As the GLM focused on individual bases or base pairs in the Alu consensus sequence, we also developed our own k-mer based method to search for sequence motifs that influence Alu expression, better suited to the data at hand than regular motif search. First, to circumvent the problems that arise from the use of a reference sequence, we constructed a de Bruijn graph from the k-mers of all Alu sequences ($k = 9$, for the tuning of k and its justification, see Methods). Next, we filtered for k-mers with statistical signifi-

Table 1 De Bruijn graph significant k-mers

Shown are the k-mers ($k = 9$) from the de Bruijn graph created of all Alu sequences (incl. 100 bp flanking regions) that passed all significance and relevance thresholds: $p \leq 0.05$ (Fisher test with Bonferroni multiple testing correction) and $OR \leq 0.5$ or $OR \geq 2.0$ for both test schemes. Listed are the k-mer sequence, associated OR (k-mer occurrence versus read counts), JASPAR match, and associated Matrix ID (see Supplement S10 for expanded statistics). The binding profile matches described in the Results are marked as bold.

	k-mer	OR	JASPAR match	Matrix ID
1	AACGCGCCA	2.65	—	
2	ATCGCCCGC	2.72	**NFIX**	MA0671.1
			NR2C2 (var.2)	MA1536.1
3	CGGACTGCT	2.07	MEIS1	MA0498.2
			TEAD3	MA0808.1
4	CTCAACGCC	2.40	SOX18	MA1563.1
			BARHL1	MA0877.2
			ZNF354C	MA0130.1
			NR2C2 (var.2)	MA1536.1
			GSX2	MA0893.2
5	GAAACCGTC	2.12	—	
6	GACACGCGC	2.27	**ARNT::HIF1A**	MA0259.1
			TFE3	MA0831.2
			USF1	MA0093.1
7	GATCGCCCG	2.56	GATA2	MA0036.1
8	GGCGGACTG	2.35	MEIS1	MA0498.2
9	GGGCGGACT	2.64	—	
10	TAGGCGCGC	2.08	—	
11	TCAACGCCT	2.22	TBX4	MA0806.1
			TBX5	MA0807.1
			NR2C2 (var.2)	MA1536.1
			GSX2	MA0893.2
			MGA	MA0801.1
12	TGACACGCG	2.92	FOS::JUN	MA0099.2
			TFE2	MA0831.2
			TBX4	MA0806.1
			MGA	MA0801.1
			TBX5	MA0807.1
			USF1	MA0093.1

cance and biological relevance. Each k-mer splits the group of all Alu sequences into those containing the k-mer and those that do not. At the same time, each Alu element can be either expressed (at least one read count) or not. This gives rise to a 2×2 contingency table, to which we apply a Fisher test and calculate the odds ratio (OR, see Methods). k-mers with both an OR less than 0.5 or greater than 2 and a Bonferroni-corrected Fisher test p-value less than 0.05 were included in the downstream analysis, leaving us with 38 k-mers of interest. Finally, we performed another Fisher test comparing expressed and non-expressed Alu elements in the group containing a specific k-mer of interest versus the group of Alu elements containing the k-1 Prefix and suffix of that k-mer. This resulted in merely 12 remaining k-mers of interest after applying the same stringency criteria (Table 1). We also ensured that the reported k-mers originate from the Alu sequence itself and not from the 100 bp flanking region (see Methods and Supplement S11).

A comparison of those k-mers with transcription factor binding profiles from the JASPAR database reveals that 8 of the 12 k-mers show similarities to known motifs (Fornes *et al.* 2020). The most intriguing findings to us were the k-mers resembling the binding profiles of ARNT::HIF1A (MA0259.1, which is a master transcriptional regulator of hypoxia response, NFIX (MA0671.1, which is involved in the replication of adenovirus 2, and TFE3 (MA0831.2, which plays a role in the general immune response. Also, all listed transcription factors are Pol-II specific (see Supplement S10 for references and detailed statistics).

DISCUSSION

We presented a multifaceted analysis of the Alu elements found in the human genome in K562 cells, focusing on their RNA metabolism and sequence features that influence it.

The read count distributions of Alu elements in comparison to mRNA presented in Figure 2a, b, and e correspond well to previous *in vitro* findings, corroborating that Alu expression is generally low in comparison to gene transcription and that the younger Alu families (AluY and subfamilies) are expressed at even lower rates than the older Alu families (AluJ and AluS) (Paulson and Schmid 1986; Bennett *et al.* 2008).

In contrast though, the results of our RNA half-life estimation of Alu transcripts through the use of TT-seq (Schwalb *et al.* 2016) and 4sU-seq (Miller *et al.* 2011) are surprising. Previous studies suggest that Alu RNAs should be less stable than regular gene transcripts, as they contain Adenine and Uracil rich element (ARE) motifs, a sequence feature linked to decreased transcript stability with an as yet unknown mechanism (An *et al.* 2004). In contrast, our MLE predicts that Alu transcripts and mRNAs should be of similar stability (Figure 2c and d). However, our MLE can only serve as an assessment to compare the relative half-life distributions; its predictions do not represent explicit half-life values. Firstly, the MLE has to contend with the overall paucity of Alu read counts, which limits its explanatory power, an issue that could only be remedied by vastly increasing sequencing depths in future experiments. Secondly, to not exacerbate the already demanding estimation, our model contains some simplifications, such as the assumption of steady-state conditions, using a Poisson distribution to model read counts instead of a zero-inflated negative binomial distribution, and neglecting non-constant labeling efficiencies for short labeling periods (see Supplement S5). As such, our half-life estimates point towards a need for further research regarding the stability of Alu transcripts, as the subject matter is less straightforward as it appears prima facie. Also, our data originates from K562 cells and is thus subject to the peculiarities of that cell line (Li *et al.* 2000). Nonetheless, taking into account the nonautonomous life cycle of Alu elements and their reliance on LINE-1 repeat translation (Boeke 1997; Dewannieux *et al.* 2003) in conjunction with their low expression, transcript persistence in the cell could potentially be instrumental for successful retrotransposition.

Regarding our non-interventional analysis concerning the origin of Alu transcripts, we observed a higher expression of intragenic Alu elements compared to intergenic ones (Figure 3a). This on its own cannot be taken as solid evidence either for or against a connection between Alu and gene transcription. Genome accessibility is a confounding factor, as intragenic regions are in general more accessible than intergenic ones (Guo *et al.* 2017). It should also be noted that a screening of Alu elements proximal to Pol-I and Pol-III transcripts, while intriguing, falls outside the scope of our study, as our approach exploits the large group size of intra- and intergenic Alu elements to compensate for the low read counts. The difference in expression between intragenic Alu elements that lie in sense or antisense direction with respect to their associated gene is so minor that we deem it irrelevant (Figure 3b). Therefore, we analyzed the correlation between Alu and gene transcription (Figure 3c). A certain level of correlation is to be expected regardless of Alu transcript origin due to genome accessibility, as mentioned above. However, the lack of a difference in correlation strength between sense and antisense points towards Alu transcripts not originating primarily as side products of Pol-II gene transcription. If that were the case, a stronger correlation between Alu elements in sense direction with respect to their associated gene and the expression of that gene should exist. This is further corroborated by the number of sense and antisense Alu elements present in each family (Figure 3d). If Alu transcripts were primarily side products of Pol-II gene transcription, Alu insertion should show a bias towards aligning new Alu elements in sense direction with respect to their associated gene. Taken together, we interpret these findings as weakening the hypothesis that Alu transcripts are mainly created alongside Pol-II gene transcription. This does not rule out the possibility for a fraction of Alu elements being transcribed in this fashion, but it does not appear to be a major contributor to Alu expression in the cell.

Our analysis of Pol-II and Pol-III occupancy data did not detect any correlation between Alu expression and polymerase occupancy (Supplement S7b), which was a likely outcome given the nonconformity of the two measurements (Ehrensberger *et al.* 2013; Gressel *et al.* 2017). The issue of such a correlation was also addressed recently by Zhang *et al.* (2019), who utilized an extensive collection of data produced as part of the ENCODE project in conjunction with RAMPAGE (Batut and Gingeras 2013). While their study covered only 1.5% of all annotated Alu elements in the human genome due to strict read-mapping and analysis constraints, they identified a subset of Alu elements as Pol-III transcribed, also in accordance with chromatin immunoprecipitation (ChIP)-seq data. This suggests that the detection of correlation between Alu expression and polymerase occupancy, while possible, necessitates high-resolution data that makes a global assessment demanding and beyond the explanatory power to be found in our data.

The differential expression analysis under α-amanitin inhibition of Pol-II suggests that a substantial part of Alu transcription is Pol-II dependent (Figure 3e and f). To the author's knowledge, this is the first direct experimental evidence for the origin of Alu transcripts. If Alu transcripts would originate mainly from Pol-III activity, no downregulation should be present. While this may appear to conflict with our previous findings that Alu expression

is unlikely to be primarily a side product of Pol-II gene transcription, this leaves the possibility open that Alu transcripts arise from Pol-II activity independent from gene transcription. However, it seems more likely that Alu RNAs arise from different modes of transcription simultaneously, as also suggested by Conti *et al.* (2015) and Zhang *et al.* (2019). This is of particular interest regarding differential expression analyses of the past that utilized Pol-II inhibition, as it is common practice to rely on SINEs as a non-Pol-II-dependent negative control group in such experiments. Alus are the most common SINE in the human genome (Cordaux and Batzer 2009), which may call the results of these studies into question. Regarding our own results, the small but not negligible fraction of seemingly upregulated Alu and mRNA transcripts may be explained by inhibition of nuclear export (Bahar Halpern *et al.* 2015; Zander *et al.* 2016). To achieve noticeable differential expression of Alu elements under α-amanitin inhibition of Pol-II, an incubation time of 8 h was required (see Methods). This exposes the cells to extreme levels of stress, as their protein metabolism is breaking down. This, in turn, may lead to the retention and accumulation of RNAs that are not related to shock response in the nucleus. As our DTA does not differentiate between nuclear and cytoplasmic RNA fractions, this could appear as apparent upregulation. Alternatively, protein translation inhibition, especially of proteins related to mRNA degradation (such as deadenylases or Xrn1) could be involved (Orban and Izaurralde 2005; Dobin *et al.* 2016). Changes in protein level, even after extended exposure, seem to be an uncommon effect of α-amanitin treatment, according to Fiume and Stirpe (1966) and Stirpe and Fiume (1967). A reduction in protein abundance or concentration would approximately lead to a proportional decrease in degradation rates. Hence it would merely lead to a proportional shift in total expression levels. The reduction in the protein concentration might therefore lead to compensatory effects, which would make a quantitative, but not a qualitative change. Regarding Figure 3e and f is should also be noted, that an average l2fc of -0.35 ($\sigma \approx 0.99$) for mRNAs may appear small when considered individually, especially in the 2D density heatmap, where strongly expressed genes do appear to be affected more, as random measurement noise is visually overrepresented and optically dominates the plot. However, calculating the overall differential expression shows a global reduction in mRNA levels by a factor of 3 (l2fc of -1.64), comparable to, for example, the effect of heat shock response (Mahat *et al.* 2016).

One influential position resulting from our GLM analysis of all Alu sequences is of particular note (see Figure 5). Position 79 lies within the B box of the Pol-III promoter sequence and is most sensitive to insertions, which bring it closer to the consensus sequence of the tRNA promoter: GWTCRANNC (Paolella *et al.* 1983). This suggests that the promoter plays a role in Alu transcription, further substantiating the assumption that Alu RNAs arise from various methods of transcription.

It is remarkable that our new method employing de Bruijn graph k-mers does detect statistically significant and biologically relevant sequence motifs when previous attempts using regular *de novo* motif search did not (Zhang *et al.* 2019). We attribute this to some extent to the tuning of the parameter k (see Methods). We found that with a k either smaller or larger than 9, the high sequence similarity of Alu elements quickly leads to problems. With a k smaller than 9, many k-mers are possessed by all or almost all Alu elements, while with a k larger than 9, a high number of k-mers are unique to a few or even a single Alu sequence. In both unfavorable cases, a majority of k-mers are ruled out simply by the selection of k. This might appear as a limitation of our method, but in fact, it is an intrinsic limitation shared by all continuous-motif-based methods. Our method merely makes this problem noticeable and offers a solution through parameter tuning. We also see the potential for our approach to be applied to other data sets in the future.

Taken together, we present a compendium of results regarding the RNA metabolism of Alu elements. Our analyses affirm that older Alu families are more strongly expressed than younger Alu families, but the stability of Alu transcripts appears to be close to that of mRNA. Furthermore, we find evidence for Alu transcripts originating both from Pol-II and Pol-III activity, but no evidence for Alu transcripts being side products of Pol-II gene transcription. Finally, we have identified a list of sequence features that influence Alu transcription and are thus targets for further investigations, and developed a novel method to test the statistical significance and biological relevance of de Bruijn graph k-mers.

Data Availability

The sequencing data used in this publication have been deposited in NCBI's Gene Expression Omnibus (Edgar *et al.* 2002) and are accessible through GEO Series accession number GSE75792 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE75792) for the data described in Schwalb *et al.* (2016), and through GEO Series accession number GSE185485 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE185485) for the data relating to the inhibition of RNA Pol-II by α-amanitin. GSE75792 was used for all non-interventional analyses, while GSE185485 was used for the differential expression analysis.

LITERATURE CITED

Ade, C., A. M. Roy-Engel, and P. L. Deininger, 2013 Alu elements: An intrinsic source of human genome instability.

An, H. J., D. Lee, K. H. Lee, and J. Bhak, 2004 The association of Alu repeats with the generation of potential AU-rich elements (ARE) at 3′ untranslated regions. BMC Genomics **5**: 1–5.

Bahar Halpern, K., I. Caspi, D. Lemze, M. Levy, S. Landen, *et al.*, 2015 Nuclear Retention of mRNA in Mammalian Tissues. Cell Reports **13**: 2653–2662.

Bateman, A., M. J. Martin, S. Orchard, M. Magrane, R. Agivetova, *et al.*, 2021 UniProt: The universal protein knowledgebase in 2021. Nucleic Acids Research **49**: D480–D489.

Batut, P. and T. R. Gingeras, 2013 RAMPAGE: Promoter activity profiling by paired-end sequencing of 5′-complete cDNAs. Current Protocols in Molecular Biology **104**: 25B.11.1–25B.11.16.

Batzer, M. A., P. L. Deininger, U. Hellmann-Blumberg, J. Jurka, D. Labuda, *et al.*, 1996 Standardized nomenclature for Alu repeats. In *Journal of Molecular Evolution*, volume 42, pp. 3–6.

Bennett, E. A., H. Keller, R. E. Mills, S. Schmidt, J. V. Moran, *et al.*, 2008 Active Alu retrotransposons in the human genome. Genome Research **18**: 1875–1883.

Boeke, J. D., 1997 Lines and alus — the polya connection. Nature Genetics **16**: 6–7.

Chen, L. L. and L. Yang, 2017 ALUternative Regulation for Gene Expression.

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

Cordaux, R. and M. A. Batzer, 2009 The impact of retrotransposons on human genome evolution.

Dagan, T., R. Sorek, E. Sharon, G. Ast, and D. Graur, 2004 AluGene: a database of Alu elements incorporated within protein-coding genes. Nucleic acids research **32**: D489–92.
Davis, C. A., B. C. Hitz, C. A. Sloan, E. T. Chan, J. M. Davidson, *et al.*, 2018 The Encyclopedia of DNA elements (ENCODE): Data portal update. Nucleic Acids Research **46**: D794–D801.
De Mori, R., M. Severino, M. M. Mancardi, D. Anello, S. Tardivo, *et al.*, 2019 Agenesis of the putamen and globus pallidus caused by recessive mutations in the homeobox gene GSX2. Brain **142**: 2965–2978.
Deininger, P., E. Lander, L. Linton, B. Birren, C. Nusbaum, *et al.*, 2011 Alu elements: know the SINEs. Genome Biology **12**: 236.
Deininger, P. L. and M. A. Batzer, 1999 Alu repeats and human disease. Molecular Genetics and Metabolism **67**: 183–193.
Deininger, P. L., M. A. Batzer, C. A. Hutchison, and M. H. Edgell, 1992 Master genes in mammalian repetitive DNA amplification.
Dewannieux, M., C. Esnault, and T. Heidmann, 2003 LINE-mediated retrotransposition of marked Alu sequences. Nature Genetics **35**: 41–48.
Dobin, A., T. R. Gingeras, C. Spring, R. Flores, J. Sampson, *et al.*, 2016 Mapping RNA-seq with STAR. Curr Protoc Bioinformatics **51**: 586–597.
Dunham, I., A. Kundaje, S. F. Aldred, P. J. Collins, C. A. Davis, *et al.*, 2012 An integrated encyclopedia of DNA elements in the human genome. Nature **489**: 57–74.
Edgar, R., M. Domrachev, and A. E. Lash, 2002 Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Research **30**: 207–210.
Ehrensberger, A. H., G. P. Kelly, and J. Q. Svejstrup, 2013 Mechanistic interpretation of promoter-proximal peaks and RNAPII density maps. Cell **154**: 713–715.
Evgen'ev, M. B., 2007 Mobile elements and genome evolution. Molecular Biology **41**: 203–213.
Fiume, L. and F. Stirpe, 1966 Decreased RNA content in mouse liver nuclei after intoxication with α-amanitin. BBA Section Nucleic Acids And Protein Synthesis **123**: 643–645.
Flury, B. D., 1990 Acceptance-rejection sampling made easy. SIAM Review **32**: 474–476.
Fornes, O., J. A. Castro-Mondragon, A. Khan, R. Van Der Lee, X. Zhang, *et al.*, 2020 JASPAR 2020: Update of the open-Access database of transcription factor binding profiles. Nucleic Acids Research **48**: D87–D92.
Frankish, A., M. Diekhans, I. Jungreis, J. Lagarde, J. E. Loveland, *et al.*, 2021 GENCODE 2021. Nucleic Acids Research **49**: D916–D923.
Friedman, J., T. Hastie, and R. Tibshirani, 2010 Regularization Paths for Generalized Linear Models via Coordinate Descent. Journal of Statistical Software **33**: 1–22.
Gaudet, P., M. S. Livstone, S. E. Lewis, and P. D. Thomas, 2011 Phylogenetic-based propagation of functional annotations within the Gene Ontology consortium. Briefings in Bioinformatics **12**: 449–462.
Gentles, A. J., O. Kohany, and J. Jurka, 2005 Evolutionary diversity and potential recombinogenic role of integration targets of non-LTR retrotransposons. Molecular Biology and Evolution **22**: 1983–1991.
Gressel, S., K. Lidschreiber, and P. Cramer, 2019a Transient transcriptome sequencing : experimental protocol to monitor genome-wide RNA synthesis including enhancer transcription. protocols.io pp. 1–20.
Gressel, S., B. Schwalb, and P. Cramer, 2019b The pause-initiation limit restricts transcription activation in human cells. Nature Communications **10**: 1–12.
Gressel, S., B. Schwalb, T. M. Decker, W. Qin, H. Leonhardt, *et al.*, 2017 CDK9-dependent RNA polymerase II pausing controls transcription initiation. eLife **6**: 1–24.
Guo, H., B. Hu, L. Yan, J. Yong, Y. Wu, *et al.*, 2017 DNA methylation and chromatin accessibility profiling of mouse and human fetal germ cells. Cell Research **27**: 165–183.
Han, K., M. K. Konkel, J. Xing, H. Wang, J. Lee, *et al.*, 2007 Mobile DNA in Old World monkeys: A glimpse through the rhesus macaque genome. Science **316**: 238–240.
Häsler, J. and K. Strub, 2006 Alu elements as regulators of gene expression. Nucleic Acids Research **34**: 5491–5497.
Hoeth, M., H. Niederleithner, R. Hofer-Warbinek, M. Bilban, H. Mayer, *et al.*, 2012 The Transcription Factor SOX18 Regulates the Expression of Matrix Metalloproteinase 7 and Guidance Molecules in Human Endothelial Cells. PLOS ONE **7**: e30982.
Holley, G. and P. Melsted, 2020 Bifrost: highly parallel construction and indexing of colored and compacted de Bruijn graphs. Genome biology **21**: 249.
Howe, K. L., P. Achuthan, J. Allen, J. Allen, J. Alvarez-Jarreta, *et al.*, 2021 Ensembl 2021. Nucleic Acids Research **49**: D884–D891.
Hu, X., J. Yuan, Y. Shi, J. Lu, B. Liu, *et al.*, 2012 pIRS: Profile-based illumina pair-end reads simulator. Bioinformatics **28**: 1533–1535.
Huang, X., L. Ding, K. L. Bennewith, R. T. Tong, S. M. Welford, *et al.*, 2009 Hypoxia-Inducible mir-210 Regulates Normoxic Gene Expression Involved in Tumor Initiation. Molecular Cell **35**: 856–867.
Idury, R. M. and M. S. Waterman, 1995 A new algorithm for DNA sequence assembly. Journal of computational biology : a journal of computational molecular cell biology **2**: 291–306.
Iqbal, Z., M. Caccamo, I. Turner, P. Flicek, and G. McVean, 2012 De novo assembly and genotyping of variants using colored de Bruijn graphs. Nature Genetics **44**: 226–232.
Jacob, S. T., W. Muecke, E. M. Sajdel, and H. N. Munro, 1970 Evidence for extranucleolar control of RNA synthesis in the nucleolus. Biochemical and Biophysical Research Communications **40**: 334–342.
Jagadeeswaran, P., B. G. Forget, and S. M. Weissman, 1981 Short interspersed repetitive DNA elements in eucaryotes: Transposable DNA elements generated by reverse transcription of RNA pol III transcripts?
Jurka, J. and T. Smith, 1988 A fundamental division in the Alu family of repeated sequences. Proceedings of the National Academy of Sciences of the United States of America **85**: 4775–4778.
Karolchik, D., A. S. Hinricks, T. S. Furey, K. M. Roskin, C. W. Sugnet, *et al.*, 2004 The UCSC table browser data retrieval tool. Nucleic Acids Research **32**.
Kazazian, H. H., 2004 Mobile Elements: Drivers of Genome Evolution.
Kedinger, C., M. Gniazdowski, J. L. J. Mandel, F. Gissinger, and P. Chambon, 1970 Alpha-amanitin: a specific inhibitor of one of two DNA-pendent RNA polymerase activities from calf thymus. Biochemical and biophysical research communications **38**: 165–171.
Kodeboyina, S., P. Balamurugan, L. Liu, and B. S. Pace, 2010 cJun modulates Gγ-globin gene expression via an upstream cAMP response element. Blood Cells, Molecules, and Diseases **44**: 7–15.
Kriegs, J. O., G. Churakov, J. Jurka, J. Brosius, and J. Schmitz, 2007 Evolutionary history of 7SL RNA-derived SINEs in Supraprimates.

Lander, E. S., L. M. Linton, B. Birren, C. Nusbaum, M. C. Zody, *et al.*, 2001 Initial sequencing and analysis of the human genome. Nature **409**: 860–921.
Li, T. H., C. Kim, C. M. Rubin, and C. W. Schmid, 2000 K562 cells implicate increased chromatin accessibility in Alu transcriptional activation. Nucleic Acids Research **28**: 3031–3039.
Lindell, T. J., F. Weinberg, P. W. Morris, R. G. Roeder, and W. J. Rutter, 1970 Specific Inhibition of Nuclear RNA Polymerase II by α-Amanitin. Science **170**: 447–449.
Livak, K. J. and T. D. Schmittgen, 2001 Analysis of Relative Gene Expression Data Using Real-Time Quantitative PCR and the 2-ΔΔCT Method. Methods **25**: 402–408.
Love, M. I., W. Huber, and S. Anders, 2014 Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome biology **15**: 550.
Mahat, D. B., H. H. Salamanca, F. M. Duarte, C. G. Danko, and J. T. Lis, 2016 Mammalian Heat Shock Response and Mechanisms Underlying Its Genome-wide Transcriptional Regulation. Molecular Cell **62**: 63–78.
Mariner, P. D., R. D. Walters, C. A. Espinoza, L. F. Drullinger, S. D. Wagner, *et al.*, 2008 Human Alu RNA Is a Modular Transacting Repressor of mRNA Transcription during Heat Shock. Molecular Cell **29**: 499–509.
Menon, I. A., 1971 Differential Effects of α-Amanitin on RNA Polymerase Activity in Nuclei and Mitochondria. Canadian Journal of Biochemistry **49**: 1395–1398.
Miller, C., B. Schwalb, K. Maier, D. Schulz, S. Dümcke, *et al.*, 2011 Dynamic transcriptome analysis measures rates of mRNA synthesis and decay in yeast. Molecular Systems Biology **7**: 458–458.
Nelder, J. A. and R. W. M. Wedderburn, 1972 Generalized Linear Models. Journal of the Royal Statistical Society. Series A (General) **135**: 370.
Nguyen, V. T., F. Giannoni, M.-F. Dubois, S.-J. Seo, M. Vigneron, *et al.*, 1996 In Vivo Degradation of RNA Polymerase II Largest Subunit Triggered by α-Amanitin. Nucleic Acids Research **24**: 2924–2929.
Oler, A. J., S. Traina-Dorge, R. S. Derbes, D. Canella, B. R. Cairns, *et al.*, 2012 Alu expression in human cell lines and their retrotranspositional potential. Mobile DNA **3**.
Orban, T. I. and E. Izaurralde, 2005 Decay of mRNAs targeted by RISC requires XRN1, the Ski complex, and the exosome. RNA **11**: 459–469.
Orioli, A., C. Pascali, A. Pagano, M. Teichmann, and G. Dieci, 2012 RNA polymerase III transcription control elements: Themes and variations.
Panning, B. and J. R. Smiley, 1993 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 **13**: 3231–3244.
Paolella, G., M. A. Lucero, M. H. Murphy, and F. E. Baralle, 1983 The Alu family repeat promoter has a tRNA-like bipartite structure. The EMBO Journal **2**: 691–696.
Paulson, K. E. and C. W. Schmid, 1986 Transcriptional inactivity of alu repeats in HeLa cells. Nucleic Acids Research **14**: 6145–6158.
Pevzner, P. A., H. Tang, and M. S. Waterman, 2001 An Eulerian path approach to DNA fragment assembly. Proceedings of the National Academy of Sciences of the United States of America **98**: 9748–9753.
Price, A. L., E. Eskin, and P. A. Pevzner, 2004 Whole-genome analysis of Alu repeat elements reveals complex evolutionary history. Genome Research **14**: 2245–2252.
Quentin, Y., 1992 Fusion of a free left alu monomer and a free right alu monometer at the origin of the alu family in the primate genomes. Nucleic Acids Research **20**: 487–493.
Reid, B. D. and P. Parsons, 1971 Partial Purification of Mitochondrial RNA Polymerase from Rat Liver. Proceedings of the National Academy of Sciences **68**: 2830–2834.
Richard Shen, M., M. A. Batzer, and P. L. Deininger, 1991 Evolution of the master Alu gene(s). Journal of Molecular Evolution **33**: 311–320.
Rozen, S. and H. Skaletsky, 2000 Primer3 on the WWW for general users and for biologist programmers. Methods in molecular biology (Clifton, N.J.) **132**: 365–86.
Saccone, C., R. Gallerani, M. Gadaleta, and M. Greco, 1971 The effect of α-amanitin on RNA synthesis in rat liver mitochondria. FEBS Letters **18**: 339–341.
Schmid, C. W. and P. L. Deininger, 1975 Sequence organization of the human genome. Cell **6**: 345–358.
Schwalb, B., M. Michel, B. Zacher, K. F. Hauf, C. Demel, *et al.*, 2016 TT-seq maps the human transient transcriptome. Science **352**: 1225–1228.
Schwalb, B., D. Schulz, M. Sun, B. Zacher, S. Dümcke, *et al.*, 2012 Measurement of genome-wide RNA synthesis and decay rates with Dynamic Transcriptome Analysis (DTA). Bioinformatics **28**: 884–885.
Sexton, C. E. and M. V. Han, 2019 Paired-end mappability of transposable elements in the human genome. Mobile DNA **10**: 1–11.
Sinnett, D., C. Richer, J. Deragon, and D. Labuda, 1991 Alu RNA secondary structure consists of two independent 7 SL RNA-like folding units. J. Biol. Chem. **266**: 8675–8678.
Stirpe, F. and L. Fiume, 1967 Studies on the pathogenesis of liver necrosis by alpha-amanitin. Effect of alpha-amanitin on ribonucleic acid synthesis and on ribonucleic acid polymerase in mouse liver nuclei. The Biochemical journal **105**: 779–782.
Tisdale, S. and L. Pellizzoni, 2017 RNA-Processing Dysfunction in Spinal Muscular Atrophy. In *Spinal Muscular Atrophy: Disease Mechanisms and Therapy*, pp. 113–131, Academic Press.
Ullu, E. and C. Tschudi, 1984 Alu sequences are processed 7SL RNA genes. Nature **312**: 171–172.
Weichenrieder, O., K. Wild, K. Strub, and S. Cusack, 2000 Structure and assembly of the Alu domain of the mammalian signal recognition particle. Nature **408**: 167–173.
White, R. J., 1997 Regulation of RNA polymerases i and III by the retinoblastoma protein: A mechanism for growth control?
Xing, J., Y. Zhang, K. Han, A. H. Salem, S. K. Sen, *et al.*, 2009 Mobile elements create structural variation: Analysis of a complete human genome. Genome Research **19**: 1516–1526.
Yi, C. H., A. Russ, and J. D. Brook, 2000 Virtual cloning and physical mapping of a human T-box gene, TBX4. Genomics **67**: 92–95.
Zacher, B., M. Michel, B. Schwalb, P. Cramer, A. Tresch, *et al.*, 2017 Accurate promoter and enhancer identification in 127 ENCODE and roadmap epigenomics cell types and tissues by GenoSTAN. PLoS ONE **12**: e0169249.
Zander, G., A. Hackmann, L. Bender, D. Becker, T. Lingner, *et al.*, 2016 MRNA quality control is bypassed for immediate export of stress-responsive transcripts. Nature **540**: 593–596.
Zhang, X. O., T. R. Gingeras, and Z. Weng, 2019 Genome-wide analysis of polymerase III–transcribed Alu elements suggests cell-type–specific enhancer function. Genome Research **29**: 1402–1414.
Zheng, H., H. Wang, and F. Azuaje, 2011 Incorporation of Ontology-driven biological knowledge into cardiovascular genomics. In *2011 Computing in Cardiology*, pp. 565–568.

ACKNOWLEDGMENT

We thank the Regional Computing Center of the University of Cologne (RRZK) for providing computing time on the DFG-funded (Funding number: INST 216 / 512 / 1FUGG) High Performance Computing (HPC) system CHEOPS as well as support.

Funding
Till Baar is partly funded by an SFB 680 grant by the Deutsche Forschungsgemeinschaft.