Table S2. Real-time reverse transcription polymerase chain reaction (RT-qPCR) primers used in this study. Primer sequences were designed using Primer3 software, following the parameters described in the **Methods** section. PCR efficiency values of primers (E) were calculated for each gene from the given slope after running standard curves.

Ref. #	Direction	Primer sequences	Length (nt)	Primer efficiency (E)	Target	Polymerase (Pol)
SG-116	forward	ACAATTCCAAATAGCGACCACATCA	25	0.97	Synthetic, ERCC-00043 („Spike-in 2“)	-
SG-117	reverse	TACCTCAACCCTTCCAGTGTCTAAG	25			
SG-118	forward	CATAAGCGGAGAAAGAGGGAATGAC	25	1.06	Synthetic, ERCC-00145 („Spike-in 5“)	-
SG-119	reverse	GCTAAATAGAGAGCATCCACACCTC	25			
MM-fw	forward	AGACTGGCATTCCCGTGATA	20	1.00	Synthetic, ERCC-00170 („Spike-in 12“)	-
MM-rev	reverse	GCTAAAACCCCTGCCTGCAA	20			
SG-104	forward	TGTCTGTCCTACTACCATGTCTGAA	25	1.00	Human, CWC22	Pol II
SG-105	reverse	TCCATATAAAGTGCCAAGGGTTCAC	25			
SG-84	forward	TCTCTGTTTGGAAGCCATTTGACTC	25	1.07	Human, PAIP1	Pol II
SG-85	reverse	AAAGCCTGCATTACTTCTCTAGCAC	25			
SG-98	forward	GGATTCTCCGTATTTGCGTCAGC	23	1.03	Human EGR1	Pol II
SG-99	reverse	GCTACCATTGACTCCCGAGGTA	22			
SG-126	forward	GTAACCCGTTGAACCCCATT	20	1.10	Human, 18S rRNA	Pol I
SG-127	reverse	CCATCCAATCGGTAGTAGCG	20			
SG-130	forward	CTCGCTTCGGCAGCACA	17	1.15	Human, U6 snRNA	Pol III
SG-131	reverse	AACGCTTCACGAATTTGCGT	20			

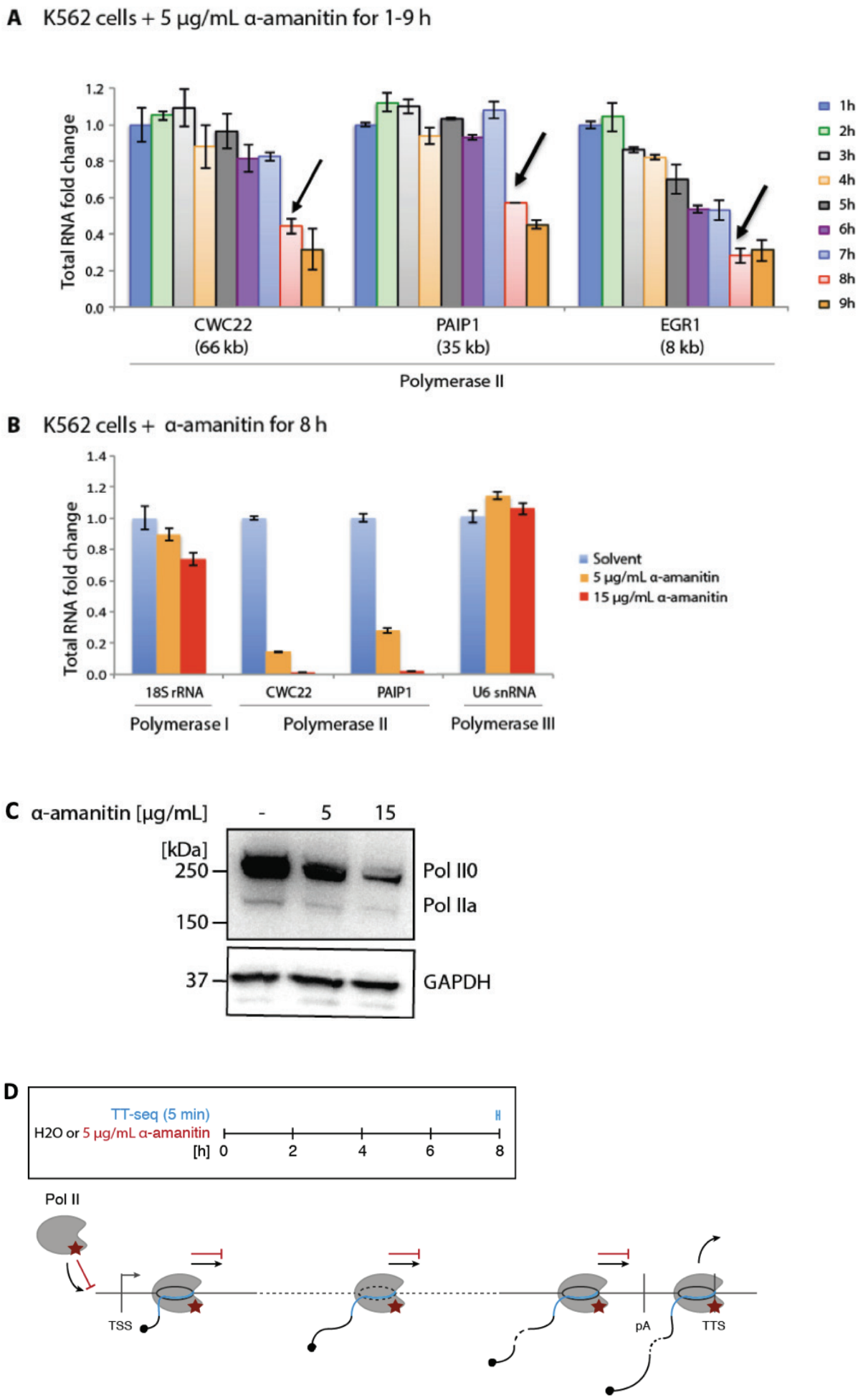


Figure S3 Supplemental figures for quality control. Treatment conditions were optimized for selective Pol-II inhibition. **a-b**) Relative expression levels of genes were analyzed using RNA spike-ins for qPCR normalization in K562 cells treated with α-amanitin (5 or 15 µg/mL) versus solvent control. The 2^(-ΔΔ 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) (Livakand Schmittgen 2001). Bars represent the means and standard errors of two technical replicates. **a**) The time point (8 h) at which about 50% decrease was observed for all Pol-II genes (CWC22, PAIP1, EGR1) is indicated by arrows. **b**) Relative expression levels of Pol-I (18S rRNA), Pol-II (CWC22, PAIP1) and Pol-III (U6 snRNA) genes were analyzed as described above. K562 cells were treated for 8 h with 5 or 15 µg/mL α-amanitin or solvent. At low α-amanitin concentrations (5 µg/mL), Pol-III and Pol-I expression levels remained close to levels in control conditions while Pol-II transcript levels decreased substantially. **c**) Degradation of the largest subunit of Pol-II, POLR2A (hRPB1) is α-amanitin dose dependent. At low α-amanitin concentrations (5 µg/mL), POLR2A levels remained close to levels in control conditions which agrees with published data (see, e.g., Fig. 1A in Nguyen et al.1996). Western blot analysis after treatment with 5 or 15 µg/mL α-amanitin for 8 h. POLR2A (bands correspond to the phosphorylated, IIo, and unphosphorylated, IIa forms of POLR2A) was visualized by N-terminal antibody F-12 (top panel). GAPDH was used as loading control (bottom panel). **d**) Graphical representation of the experimental design. TT-seq was carried out with K562 cells after treatment with solvent (water) or α-amanitin (5 µg/mL).

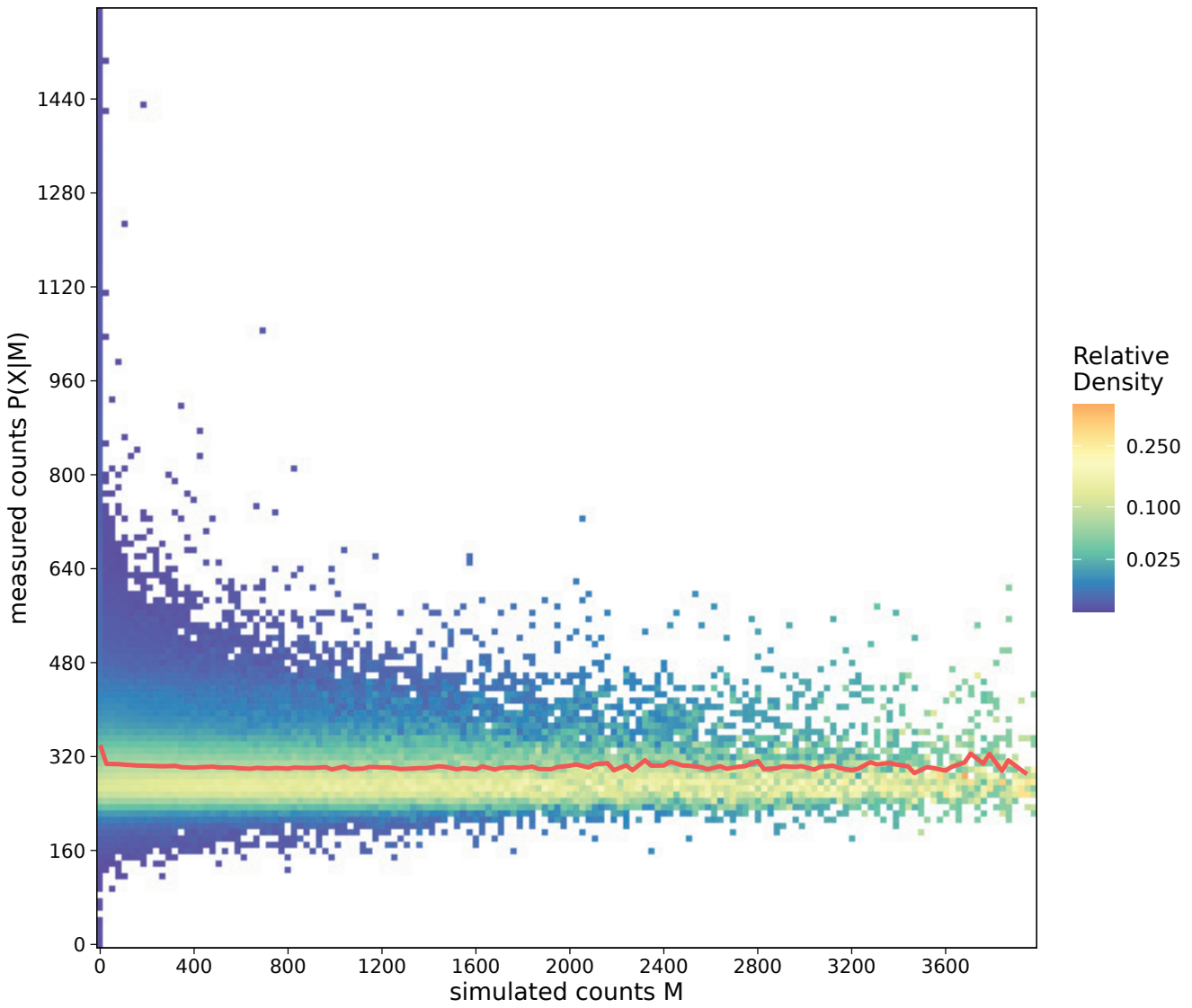


Figure S4 Alu mappability analysis — To check if mappability has a systematic effect on our measured Alu expression, we calculated a mappability score for each Alu element.

We simulated a homogenious coverage of paired-end Illumina reads using pIRS v2.0.0 (Hu 2012) with the same sequencing characteristics as those of our real sequencing experiment (see Methods). We then mapped those simulated reads back to the genome using the STAR aligner with the same settings as those used in our real sequencing experiment (see Methods and Schwalb 2016), and calculated the length corrected read counts for each Alu loci.

If there was a bias caused by mappability, the measured Alu expression would depend on this mappability. In other words, the distribution of Alu expression X, given a certain mappability score M, P(X|M), should vary. In particular, its expectation value, E(P(X|M)), should increase with increasing mappability.

The heatmap shows P(X|M) as a function of M. Each plotted column corresponds to one such distribution P(X|M) for one value of M. Relative density is color coded following the palette given on the right. The expectation value E(X|M) is indicated by the red line for each M. The figure demonstrates, that there is no systematic variation of E(X|M) with M, and that P(X|M) does not vary substantially. Thus, there is no bias caused by Alu mappability, which is also corroborated by the findings of Sexton *et al.* (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.

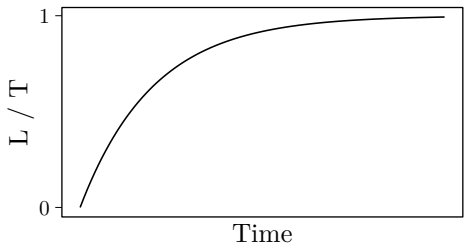
Explanation of how we come to model the degradation rate δ[1] of Alu element transcripts.

In our experiment, we measured Alu element expression using DTA (dynamic transcriptome analysis) with 4sU labeling. In this protocol, newly created transcripts are labeled, making them distinguishable from transcripts created before the labeling pulse.
In the end, we want to obtain an estimate of the half-life $t_{1/2}$ of each individual Alu element, which can be calculate from the degradation rate δ by

$$t_{1/2} = \frac{\ln(2)}{\delta}$$

The labeling pulse's duration Δt is 5 min, meaning that 5 min passed after the labeling agent was added and before the amount of labeled transcripts within the cell was measured.

With time, new transcripts are created and old transcript are degraded. This means that after the labeling pulse, the amount of unlabeled transcripts continually decreases, because all transcripts created after the labeling pulse are labeled, until only labeled transcripts remain. We assume that transcript degradation follows exponential decay, while the total amounts of transcripts in a cell remains constant, as transcription and transcript decay are in equilibrium. Therefore, the total amount of transcripts remains constant, but the ratio of labeled reads increases exponentially.



Let t_a be the total number of molecules of any Alu element a in solution. Among those t_a molecules, let l_a denote the number of newly synthesized (and therefore labeled) molecules. By assumption of steady state conditions and an exponential decay, the ratio between labeled and total molecules r_a for any Alu element a is given by

$$r_a = {}^{l_a}/_{t_a} = 1 - \exp(-\delta_a \Delta t)$$
$$\ln r_a = \ln\left(1 - \exp(-\delta_a \Delta t)\right)$$

where δ_a is the degradation rate of any Alu element a. The total amount of molecules t_a of any Alu element a is also given as:

$$t_a = {}^{\mu_a}/_{\delta_a}$$

where μ_a is the synthesis rate of any Alu element a.
Neglecting the decay of newly synthesized transcripts, we likewise obtain[2]

$$l_a = \mu_a \Delta t$$

---

[1] Usually, the degradation rate is given the symbol λ, but as λ is already in use as a parameter of the Poisson distribution, δ is used instead.

[2] This formula is an approximation. It might be replaced by the solution of the standard ODE for RNA metabolism. Or, even more difficult, it might be replaced by a term that takes into account the non-constant labeling efficiency for short labeling periods.

We further assume that

$$l_{all} = \sum_a l_a \quad \text{and} \quad t_{all} = \sum_a t_a$$

We now shift our view from the molecules in solution to read counts obtained through sequencing. We prepare a sequencing library with N_{tot} reads. After 4sU pull-down, we prepare a sequencing library of size N_{lab} reads. The distribution of total counts T_a and labeled counts L_a respectively for any Alu element a is then[3]

$$T_a \sim \text{Pois}(\lambda_{tot} = \frac{t_a}{t_{all}} \cdot N_{tot})$$
$$L_a \sim \text{Pois}(\lambda_{lab} = \frac{l_a}{l_{all}} \cdot N_{lab})$$

We thus assume that the estimated ratio between labeled and total molecules $\hat{r}_a$ for any Alu element a is given by

$$\hat{r}_a = \frac{L_a}{T_a} \cdot \frac{N_{tot}}{N_{lab}}$$

We still need to estimate ${}^{l_{all}}/_{t_{all}}$ i.e., the fraction of (all) labeled molecules among all molecules in solution. We use spike-ins to do so. Let l_{spk} be the number of labeled spike-in molecules, and t_{spk} the number of total spike-in molecules that were added to the solution. We know that

$$\frac{l_{spk}}{l_{all}} \approx \frac{L_{spk}}{L_{all}} = \frac{L_{spk}}{N_{lab}} \quad \text{and} \quad \frac{t_{spk}}{t_{all}} \approx \frac{T_{spk}}{T_{all}} = \frac{T_{spk}}{N_{tot}} \tag{1}$$

The term $q = {}^{l_{spk}}/_{t_{spk}}$ is thus known by composition of the spike-in reagents.

We use maximum likelihood estimation and the trick

$$r_a = \frac{l_a}{t_a} = \frac{l_a}{t_a} \cdot \frac{l_{spk}}{l_{spk}} \cdot \frac{t_{spk}}{t_{spk}}$$

as ${}^{l_{spk}}/_{l_{spk}}$ and ${}^{t_{spk}}/_{t_{spk}}$ are both 1, but can be rearranged to

$$= \frac{l_a}{l_{spk}} \cdot \frac{t_{spk}}{t_a} \cdot \frac{l_{spk}}{t_{spk}}$$
$$= \frac{l_a}{l_{spk}} \cdot \frac{t_{spk}}{t_a} \cdot q \tag{2}$$

We now assume that the ratio between spike-in molecules and Alu-specific molecules in solution for both fractions is approximately preserved during sequencing

$$\frac{l_a}{l_{spk}} \approx \frac{L_a}{L_{spk}} \quad \text{and} \quad \frac{t_{spk}}{t_a} \approx \frac{T_{spk}}{T_a} \tag{3}$$

Solving equation 2 suitably, we can obtain representations of both ${}^{t_a}/_{t_{all}}$ and ${}^{l_a}/_{l_{all}}$ free from any "in-solution" quantities (denoted by lower case variables), using only known "read-count"

[3] There is also the possibility to choose a (zero-inflated) negative binomial distribution instead.

quantities (denoted by upper case variables), except for q, which is also known, and r_a of course, which we want to estimate.

$$\begin{aligned} r_a &\overset{2}{=} \frac{l_a}{l_{spk}} \cdot \frac{t_{spk}}{t_a} \cdot q \\ \frac{t_a}{t_{all}} \cdot r_a &= \frac{\cancel{t_a}}{t_{all}} \cdot \frac{l_a}{l_{spk}} \cdot \frac{t_{spk}}{\cancel{t_a}} \cdot q \\ \frac{t_a}{t_{all}} &= \frac{l_a}{l_{spk}} \cdot \frac{t_{spk}}{t_{all}} \cdot \frac{1}{r_a} \cdot q \\ \frac{t_a}{t_{all}} &\overset{1,3}{\approx} \frac{L_a}{L_{spk}} \cdot \frac{T_{spk}}{N_{tot}} \cdot \frac{1}{r_a} \cdot q \end{aligned} \tag{4}$$

For ${}^{l_a}/_{l_{all}}$, we begin with this tautology, which we multiply by $r_a \,/\, ({}^{l_a}/_{l_{spk}} \cdot {}^{t_{spk}}/_{t_a} \cdot q)$, as this is 1 according to (2).

$$\begin{aligned} \frac{l_a}{l_{all}} &= \frac{l_a}{l_{all}} \\ &= \frac{l_a}{l_{all}} \cdot \left(\frac{r_a}{\frac{l_a}{l_{spk}} \cdot \frac{t_{spk}}{t_a} \cdot q} \right) \\ &= \frac{l_{spk}}{l_{all}} \cdot \frac{t_a}{t_{spk}} \cdot r_a \cdot \frac{1}{q} \\ &\overset{1,3}{\approx} \frac{L_{spk}}{N_{lab}} \cdot \frac{T_a}{T_{spk}} \cdot r_a \cdot \frac{1}{q} \end{aligned} \tag{5}$$

Using 4 and 5, we write the expected occurrence rate for total reads λ_{tot} and labeled reads λ_{lab}

$$\begin{aligned} \lambda_{tot} &= N_{tot} \cdot \frac{t_a}{t_{all}} \\ &\overset{4}{=} N_{tot} \cdot \frac{L_a}{L_{spk}} \cdot \frac{T_{spk}}{N_{tot}} \cdot \frac{1}{r_a} \cdot q \\ &= \underbrace{\frac{L_a}{L_{spk}} \cdot T_{spk} \cdot q}_{=c_{tot}} \cdot r_a^{-1} = \boxed{c_{tot} \cdot \; r_a^{-1}} \end{aligned} \tag{6}$$

$$\begin{aligned} \lambda_{lab} &= N_{lab} \cdot \frac{l_a}{l_{all}} \\ &\overset{5}{=} N_{lab} \cdot \frac{L_{spk}}{N_{lab}} \cdot \frac{T_a}{T_{spk}} \cdot r_a \cdot \frac{1}{q} \\ &= \underbrace{L_{spk} \cdot \frac{T_a}{T_{spk}} \cdot \frac{1}{q}}_{=c_{lab}} \cdot r_a = \boxed{c_{lab} \cdot \; r_a} \end{aligned} \tag{7}$$

We can now write down the likelihood function $\mathcal{L}$ of observing a specific T_a and L_a, as a parameterized combination of the individual two Poisson probability mass functions, which are given by

$$\mathrm{P}(k;\lambda) = \frac{\lambda^k \cdot \mathrm{e}^{-k}}{k!} = \mathrm{e}^{-\lambda} \cdot \frac{\lambda^k}{k!}$$

where k is number of occurrences. In our case, this is T_a and L_a respectively.

$$\begin{aligned}\mathcal{L}(L_a, T_a; r_a, q, L_{spk}, T_{spk}) &= \text{Pois}(T_a; \lambda_{tot} = c_{tot} \cdot r_a^{-1}) \cdot \text{Pois}(L_a; \lambda_{lab} = c_{lab} \cdot r_a^{-1}) \\ &= \mathrm{e}^{-\lambda_{tot}} \frac{\lambda_{tot}^{T_a}}{T_a!} \cdot \mathrm{e}^{-\lambda_{lab}} \frac{\lambda_{lab}^{L_a}}{L_a!} \\ &= \mathrm{e}^{-\left(c_{tot} \cdot r_a^{-1}\right)} \frac{\left(c_{tot} \cdot r_a^{-1}\right)^{T_a}}{T_a!} \cdot \mathrm{e}^{-(c_{lab} \cdot r_a)} \frac{(c_{lab} \cdot r_a)^{L_a}}{L_a!}\end{aligned}$$

We now ignore all constants without direct influence on r_a, as they do not influence the maximization of $\mathcal{L}$ with regards to r_a.

$$\begin{aligned}&\propto \mathrm{e}^{-\left(c_{tot} \cdot r_a^{-1}\right)} \cdot r_a^{-T_a} \cdot \mathrm{e}^{-(c_{lab} \cdot r_a)} \cdot r_a^{L_a} \\ &= \exp\left(-\frac{c_{tot}}{r_a} - c_{lab} \cdot r_a\right) \cdot r_a^{L_a - T_a}\end{aligned}$$

To find the optimal r_a, and thereby the degradation rate δ_a and the half-life $t_{1/2}$, for any Alu element a, we maximize the log likelihood function ℓ

$$\begin{aligned}\ell(r_a; L_a, T_a, q, L_{spk}, T_{spk}) &= -\frac{c_{tot}}{r_a} - c_{lab} \cdot r_a + (L_a - T_a) \cdot \ln(r_a) \\ &\overset{6,7}{=} -\frac{\left(L_a / L_{spk} \cdot T_{spk} \cdot q\right)}{r_a} - \left(L_{spk} \cdot \frac{T_a}{T_{spk}} \cdot \frac{1}{q}\right) \cdot r_a + (L_a - T_a) \cdot \ln(r_a) \\ &= L_a \left(\ln(r_a) - \frac{T_{spk}}{L_{spk}} \frac{q}{r_a}\right) + T_a \left(-\frac{L_{spk}}{T_{spk}} \frac{r_a}{q} - \ln(r_a)\right)\end{aligned}$$

Assuming the log likelihood function is well-behaved, the maximum of ℓ lies where the partial derivative of ℓ with respect to r_a is 0.

$$\frac{\partial \ell}{\partial r_a} = \frac{L_a \cdot (L_{spk} \cdot r_a + T_{spk} \cdot q)}{L_{spk} \cdot r_a^2} - \frac{T_a \cdot (L_{spk} \cdot r_a + T_{spk} \cdot q)}{T_{spk} \cdot r_a \cdot q} \coloneqq 0$$

Under the valid assumtion that r_a, L_a, T_a, q, L_{spk}, and T_{spk} are all positive, this collapses to

$$r_a = \frac{L_a}{T_a} \cdot \frac{T_{spk}}{L_{spk}} \cdot q$$

■

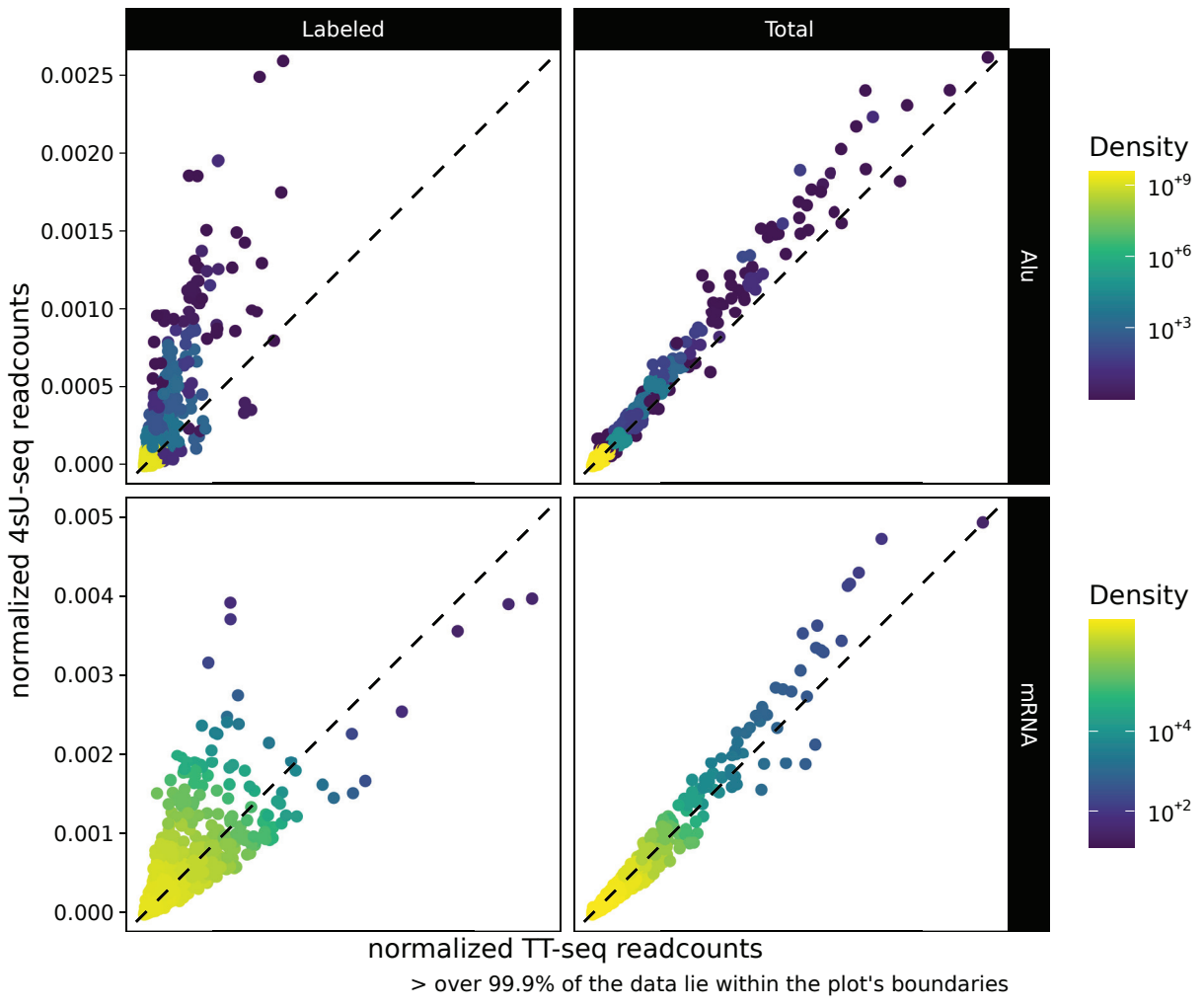


Figure S6 Correlation between sequencing methods — Scatterplot showing the correlation between the sequencing methods TT-seq (horizontal axis) and 4sU-seq (vertical axis) used for the half-life estimation via MLE. Shown are readcounts rescaled to the fraction of total reads for each group with merged replicates, excluding values of 0. The color represents point density (shown right). Both sequencing methods were used for the estimation, as a Spearman correlation r > 0.80 was observed between TT-seq and 4sU-seq readcounts in all groups. Replicates were merged, as they also showed correlation of r > 0.80 in all comparisons.

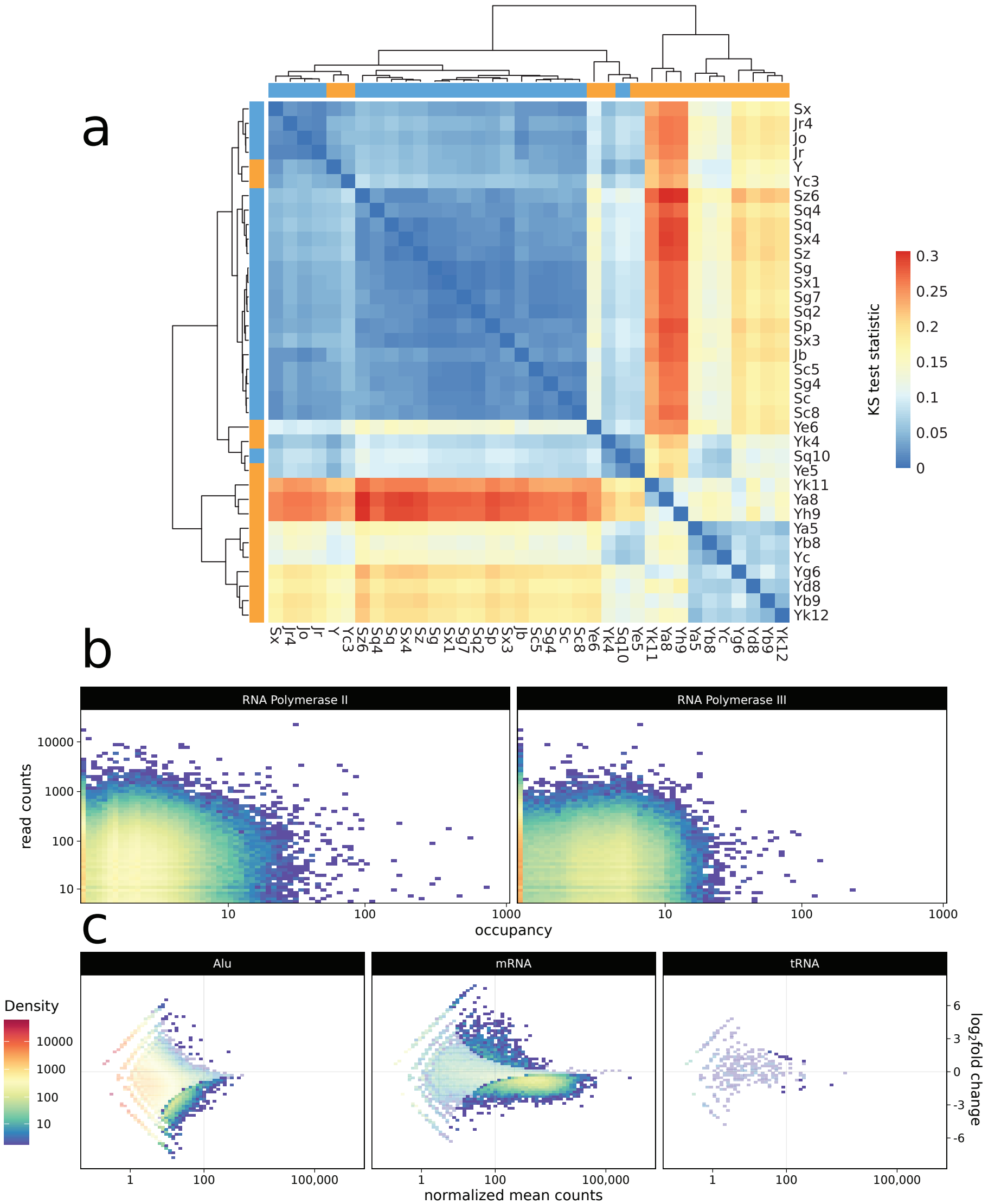


Figure S7 Supplemental figures. **a**) Heatmap of the pairwise KS test statistic between all Alu families with hierarchical clustering applied. AluY families are clearly separated from the older families, with the exception of AluSq10 and AluYc3 (AluYc2), as indicated by the colored bar (blue indicating old and yellow indicated young families). **b**) 2D density heatmap showing the lack of correlation between Alu element read counts and Pol-II (left) and Pol-III (right) occupancy. **c**) 2D density heatmap showing the DESeq2 differential expression of Alu elements, mRNAs, and tRNAs as control under a-amanitin Pol-II inhibition, using spike-in control as normalization instead of mtRNAs. Semi-transparent areas do not pass the significance threshold. Both Alu elements and mRNAs show stronger significant down- than upregulation.

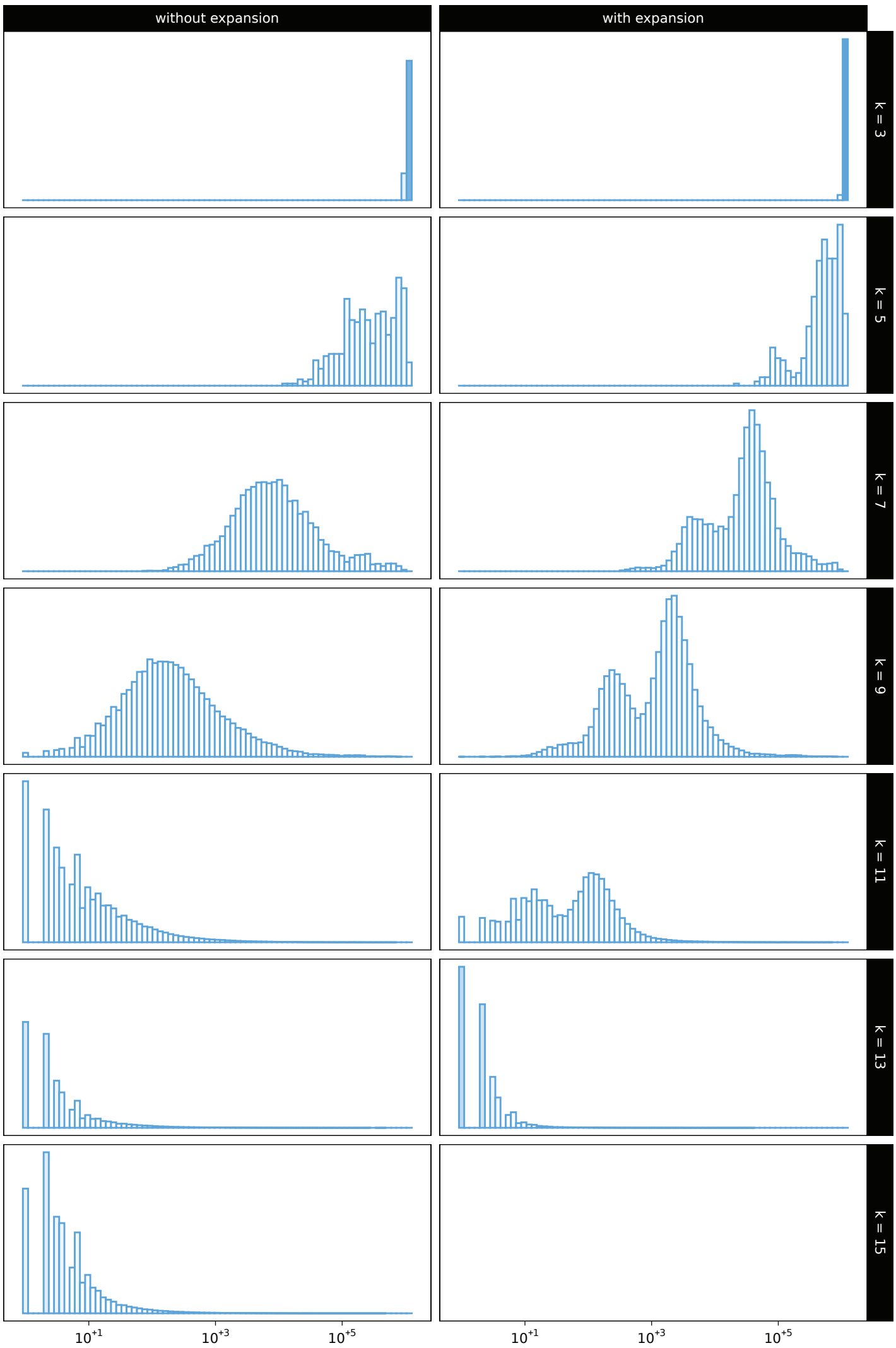


Figure S9 Choice of k — Histogram of the Alu sequences associated with each individual node in the de Bruijn graph depending on the choice of k and whether or not to expand the Alu sequences by 100 bp up- and downstream. k = 9 results in a smooth distribution of sequences per node for both cases. The bimodal distribution observable when including 100 bp up- and downstream of the Alu sequences results from the differences in sequence variability between the Alu sequence itself and the flanking regions.

Table S10 Detailed results of the JASPAR database search (Fornes 2020, relative profile score threshold of 80%) as shown in Table 1. In addition to the information reproduced directly from the JASPAR database, the column *Pol-II Reference* lists references for the Pol-II specificity of each matching transcription factor. NTNU SB: inferred from NTNU SB sequence alignment according to the UniProt database (Bateman 2021). Ensemble: Extracted from the Ensemble database (Howe 2021). BHF-UCL: Extracted from the BHF-UCL gene association of the Cardiovascular Gene Ontology Annotation Initiative (Zheng 2011).

	k-mer	OR	JASPAR match	Matrix ID	Score	Relative Score	Start	End	Strand	Predicted Sequence	Pol-II Reference
1	AACGCGCCA	2.65	–								
2	ATCGCCCGC	2.72	NFIX	MA0671.1	3.2396	0.8217	1	9	+	ATCGCCCGC	NTNU SB
			NR2C2 (var.2)	MA1536.1	0.2758	0.8182	1	8	-	GGCGTTGA	NTNU SB
3	CGGACTGCT	2.07	MEIS1	MA0498.2	1.0797	0.8036	1	7	+	CGGACTG	Gaudet 2011
			TEAD3	MA0808.1	2.5206	0.8314	1	8	-	GCAGTCCG	NTNU SB
4	CTCAACGCC	2.40	SOX18	MA1563.1	7.2890	0.9187	1	8	+	CTCAACGC	Hoeth 2012
			BARHL1	MA0877.2	4.5846	0.8548	1	8	+	CTCAACGC	NTNU SB
			ZNF354C	MA0130.1	4.6356	0.8126	1	6	+	CTCAAC	Gaudet 2011
			NR2C2 (var.2)	MA1536.1	0.5668	0.8235	1	8	-	GCGTTGAG	NTNU SB
			GSX2	MA0893.2	2.0604	0.8152	2	9	-	GGCGTTGA	DeMori 2019
5	GAAACCGTC	2.12	–								
6	GACACGCGC	2.27	ARNT::HIF1A	MA0259.1	7.7207	0.8962	2	9	-	GCGCGTGT	Huang 2009
			TFE3	MA0831.2	6.7330	0.8555	1	8	-	CGCGTGTC	Ensembl
			USF1	MA0093.1	5.7024	0.8008	2	8	-	CGCGTGT	BHF-UCL
7	GATCGCCCG	2.56	GATA2	MA0036.1	4.6275	0.8825	1	5	-	CGATC	Ensembl
8	GGCGGACTG	2.35	MEIS1	MA0498.2	1.0797	0.8036	3	9	+	CGGACTG	Gaudet 2011
9	GGGCGGACT	2.64	–								
10	TAGGCGCGC	2.08	–								
11	TCAACGCCT	2.22	TBX4	MA0806.1	5.6304	0.8382	2	9	-	AGGCGTTG	Yi 2000
			TBX5	MA0807.1	5.8659	0.8193	2	9	-	AGGCGTTG	BHF-UCL
			NR2C2 (var.2)	MA1536.1	0.2758	0.8182	1	8	-	GGCGTTGA	NTNU SB
			GSX2	MA0893.2	2.0604	0.8152	1	8	-	GGCGTTGA	DeMori 2019
			MGA	MA0801.1	3.8722	0.8126	2	9	-	AGGCGTTG	NTNU SB
12	TGACACGCG	2.92	FOS::JUN	MA0099.2	5.7234	0.8196	1	7	+	TGACACG	Kodeboyina 2010
			TFE2	MA0831.2	6.7330	0.8555	2	9	-	CGCGTGTC	BHF-UCL
			TBX4	MA0806.1	5.9236	0.8441	1	8	-	GCGTGTCA	Yi 2000
			MGA	MA0801.1	5.3019	0.8386	1	8	-	GCGTGTCA	NTNU SB
			TBX5	MA0807.1	6.0479	0.8240	1	8	-	GCGTGTCA	BHF-UCL
			USF1	MA0093.1	5.7024	0.8008	3	9	-	CGCGTGT	BHF-UCL

Bateman 2021 Bateman, A., Martin, M. J., Orchard, S., Magrane, M., Agivetova, R., Ahmad, S., Alpi, E., Bowler-Barnett, E. H., Britto, R., Bursteinas, B., Bye-A-Jee, H., Coetzee, R., Cukura, A., Silva, A. Da, Denny, P., Dogan, T., Ebenezer, T. G., Fan, J., Castro, L. G., … Zhang, J. (2021). *UniProt: The universal protein knowledgebase in 2021.* Nucleic Acids Research, 49(D1), D480–D489. https://doi.org/10.1093/nar/gkaa1100

Howe 2021 Howe, K. L., Achuthan, P., Allen, J., Allen, J., Alvarez-Jarreta, J., Amode, M. R., Armean, I. M., Azov, A. G., Bennett, R., Bhai, J., Billis, K., Boddu, S., Charkhchi, M., Cummins, C., Da Rin Fioretto, L., Davidson, C., Dodiya, K., El Houdaigui, B., Fatima, R., … Flicek, P. (2021). *Ensembl 2021.* Nucleic Acids Research, 49(D1), D884–D891. https://doi.org/10.1093/nar/gkaa942

Fornes 2020 Fornes, O., Castro-Mondragon, J. A., Khan, A., Van Der Lee, R., Zhang, X., Richmond, P. A., Modi, B. P., Correard, S., Gheorghe, M., Baranašić, D., Santana-Garcia, W., Tan, G., Chèneby, J., Ballester, B., Parcy, F., Sandelin, A., Lenhard, B., Wasserman, W. W., & Mathelier, A. (2020). *JASPAR 2020: Update of the open-Access database of transcription factor binding profiles.* Nucleic Acids Research, 48(D1), D87–D92. https://doi.org/10.1093/nar/gkz1001

DeMori 2019 De Mori, R., Severino, M., Mancardi, M. M., Anello, D., Tardivo, S., Biagini, T., Capra, V., Casella, A., Cereda, C., Copeland, B. R., Gagliardi, S., Gamucci, A., Ginevrino, M., Illi, B., Lorefice, E., Musaev, D., Stanley, V., Micalizzi, A., Gleeson, J. G., … Valente, E. M. (2019). *Agenesis of the putamen and globus pallidus caused by recessive mutations in the homeobox gene GSX2.* Brain, 142(10), 2965–2978. https://doi.org/10.1093/brain/awz247

Hoeth 2012 Hoeth, M., Niederleithner, H., Hofer-Warbinek, R., Bilban, M., Mayer, H., Resch, U., Lemberger, C., Wagner, O., Hofer, E., Petzelbauer, P., & de Martin, R. (2012). *The Transcription Factor SOX18 Regulates the Expression of Matrix Metalloproteinase 7 and Guidance Molecules in Human Endothelial Cells.* PLOS ONE, 7(1), e30982. https://doi.org/10.1371/journal.pone.0030982

Gaudet 2011 Gaudet, P., Livstone, M. S., Lewis, S. E., & Thomas, P. D. (2011). *Phylogenetic-based propagation of functional annotations within the Gene Ontology consortium.* Briefings in Bioinformatics, 12(5), 449–462. https://doi.org/10.1093/bib/bbr042

Zheng 2011 Zheng, H., Wang, H., & Azuaje, F. (2011). *Incorporation of Ontology-driven biological knowledge into cardiovascular genomics.* 2011 Computing in Cardiology, 565–568.

Kodeboyina 2010 Kodeboyina, S., Balamurugan, P., Liu, L., & Pace, B. S. (2010). *cJun modulates Gγ-globin gene expression via an upstream cAMP response element.* Blood Cells, Molecules, and Diseases, 44(1), 7–15. https://doi.org/https://doi.org/10.1016/j.bcmd.2009.10.002

Huang 2009 Huang, X., Ding, L., Bennewith, K. L., Tong, R. T., Welford, S. M., Ang, K. K., Story, M., Le, Q.-T., & Giaccia, A. J. (2009). *Hypoxia-Inducible mir-210 Regulates Normoxic Gene Expression Involved in Tumor Initiation.* Molecular Cell, 35(6), 856–867. https://doi.org/10.1016/j.molcel.2009.09.006

Yi 2000 Yi, C. H., Russ, A., & Brook, J. D. (2000). *Virtual cloning and physical mapping of a human T-box gene, TBX4.* Genomics, 67(1), 92–95. https://doi.org/10.1006/geno.2000.6222

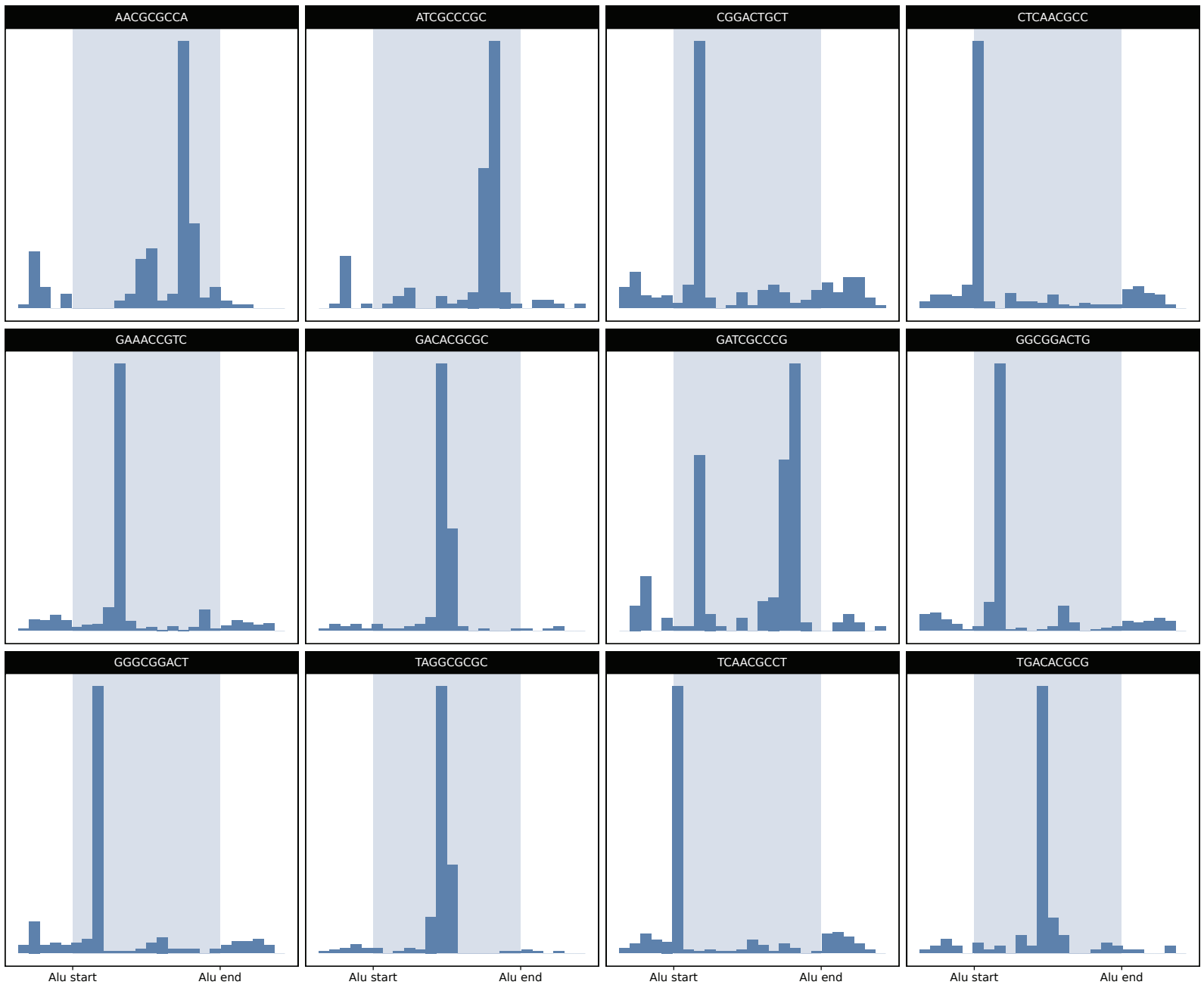


Figure S11 Relative Position of de Bruijn Graph k-mers — To make sure that the k-mers reported in Table 1 originate from the Alu sequence itself and not from the 100 bp flanking region that was included in the construction of the de Bruin graph (see Methods), we searched for each of the 12 k-mers in the sequences of all annotated Alu elements (UCSC Genome Browser).

Shown are histograms illustrating the number of k-mer sequence hits (y-axes: density) per binned relative location in all Alu elements (y-axes: relative location), relative because Alu elements vary slightly in their length. The blue area in each histogram denotes the actual Alu sequence, while the white area surrounding it left and right represents the 100 bp flanking region included in the graph construction.

For each of the 12 k-mers, the vast majority of search hits fall within the actual Alu sequence and not the flanking regions. The hits are also not situated in the variable A-rich region located in the center of the Alu sequence, but in the left or right arm. All k-mers exhibit one main location where they are found within the Alu sequence except for k-mer 7 GATCGCCCG (OR: 2.56, JASPAR hit: GATA2), which shows a secondary location of noteworthy accumulation.