Massively parallel jumping assay decodes *Alu* retrotransposition activity

Navneet Matharu[1,2,^,*], Jingjing Zhao[1,2^], Ajuni Sohota[1,2], Linbei Deng[1,2], Yan Hung[1,2] Zizheng Li[1], Jasmine Sims[1,2], Sawitree Rattanasopha[1,2], Josh Meyer[3], Lucia Carbone[3,4,5,6], Martin Kircher[7,8,*], Nadav Ahituv[1,2,*]

[1]Department of Bioengineering and Therapeutic Sciences, University of California San Francisco, San Francisco, CA, USA
[2]Institute for Human Genetics, University of California San Francisco, San Francisco, CA, USA
[3]Department of Medicine, Knight Cardiovascular Institute, Oregon Health and Science University, Portland, OR, USA
[4]Division of Genetics, Oregon National Primate Research Center, Beaverton, OR, USA
[5]Department of Molecular and Medical Genetics, Oregon Health and Science University, Portland, OR, USA
[6]Department of Medical Informatics and Clinical Epidemiology, Oregon Health and Science University, Portland, OR, USA
[7]Berlin Institute of Health of Health at Charité - Universitätsmedizin Berlin, 10178, Berlin, Germany
[8]Institute of Human Genetics, University Medical Center Schleswig-Holstein, University of Lübeck, Lübeck, Germany

^These authors contributed equally to this work
*Correspondence to: Navneet.Matharu@ucsf.edu, martin.kircher@uni-luebeck.de, nadav.ahituv@ucsf.edu

Abstract

The human genome contains millions of retrotransposons, several of which could become active due to somatic mutations having phenotypic consequences, including disease. However, it is not thoroughly understood how nucleotide changes in retrotransposons affect their jumping activity. Here, we developed a novel massively parallel jumping assay (MPJA) that can test the jumping potential of thousands of transposons *en masse*. We generated nucleotide variant library of selected four *Alu* retrotransposons containing 165,087 different haplotypes and tested them for their jumping ability using MPJA. We found 66,821 unique jumping haplotypes, allowing us to pinpoint domains and variants vital for transposition. Mapping these variants to the *Alu*-RNA secondary structure revealed stem-loop features that contribute to jumping potential. Combined, our work provides a novel high-throughput assay that assesses the ability of retrotransposons to jump and identifies nucleotide changes that have the potential to reactivate them in the human genome.

Introduction

An estimated 42% of mammalian genomes consist of retrotransposable elements that 'copy and paste' in genomes via RNA mediated transposition[1]. One of the most abundant classes of retrotransposons in mammalian genomes are '*Alu*' elements that are derived from 7SL RNA. *Alu* elements belong to the class of 'Short Interspersed Nuclear Elements' (SINEs) that have an Alu-1 restriction endonuclease site derived from *Arthrobacter Luteus* bacteria. *Alu* elements are non-autonomous retrotransposons, as they need to hijack transposition machinery from other retrotransposons (i.e., LINES/LTRs) to jump. There are roughly 1.1 million *Alu* copies in the human genome[2], which can be divided into three major subfamilies: 1) *Alu*J, the most ancient (~65 million years old, myo) and whose sequence is substantially degraded and thus believed to be inactive[3]; 2) *Alu*S, estimated to be ~30 myo and has elements that are less diverged from the consensus sequence and believed to include some active elements in the human genome; 3) *Alu*Y, the youngest lineage (~10 myo) with almost all *Alu*Y elements thought to be active. The human genome is predicted to have 852 intact functional *Alu* elements with thousands of copies that could be competent to jump[4] and it is estimated that a new *Alu* insertion happens in the human genome every 20 live births[3]. The jumping activity of *Alu* elements positively correlates with the length of their polyA tail[5,6]. Additionally, the presence of an intact 280 base pair (bp) core *Alu* sequence is thought to be essential for retrotransposition[6].

Comprehensive sequence homology analysis of the 280bp core region of 89 *Alu* elements from the human genome revealed that at least 124 nucleotide positions are conserved in all *Alu* elements that were tested to be active in retrotransposition assays[4]. However, no exhaustive functional study has been performed to date to correlate *Alu* jumping activity with sequence alterations. A high-resolution sequence-based functional analysis of the 280bp *Alu* core sequence is challenging, mainly due to lack of high-throughput retrotransposition assays. An assay for measuring the jumping activity of individual *Alu* elements has been established, utilizing splicing during the jumping event to infer antibiotic resistance, but can only test one element at a time[4]. In addition, while there have been studies that retrieve libraries of polymorphic LINE1 retrotransposons from the genome, directed evolution retrotransposition assays for *Alu* elements are lacking[7,8].

Here, we developed a massively parallel jumping assay (MPJA) that can test in parallel the ability of thousands of *Alu* elements to retrotranspose in human cells (**Fig. 1**). We used it to test the reactivation potential of three different inactive *Alu* elements and one active element. Using error prone PCR, we generated 165,087 different *Alu* haplotypes and tested their jumping ability in HeLa cells. We found 66,821 haplotypes that enabled jumping and used these results to characterize functional *Alu* domains and annotate variants that are vital for retrotransposition. We observed that even a single nucleotide change is sufficient to change the activity of an *Alu* element. Overlapping these jumping-associated variants with *Alu*-RNA secondary structure identified stem-loop features that contribute to retrotransposition potential. In summary, our results provide a novel technology that can test the jumping ability of thousands of sequences and

identifies key domains, RNA secondary structures and variants that are vital for retrotransposition.

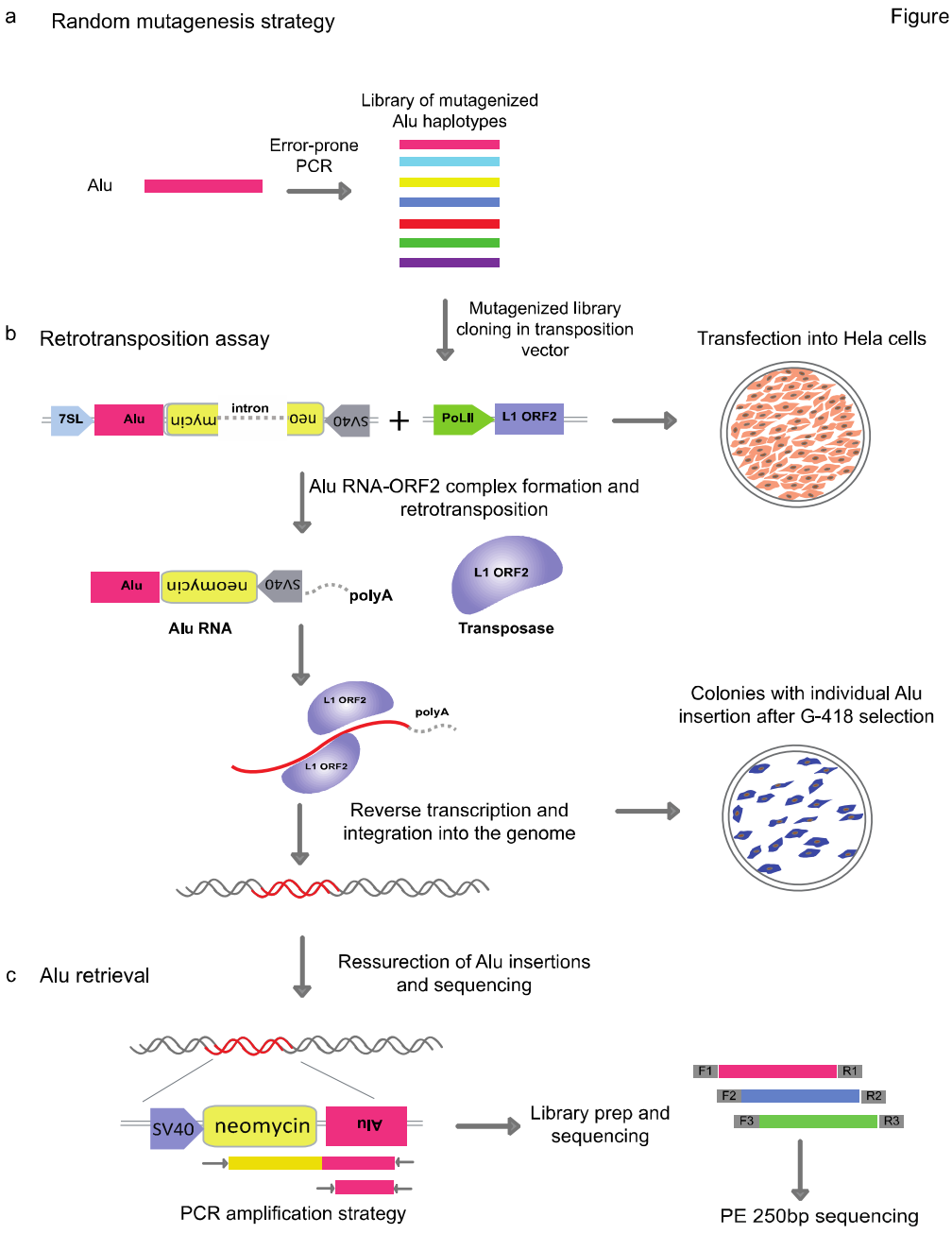

Figure 1: Massively parallel jumping assay.
a, Schematic showing the random mutagenesis strategy to generate *Alu* variant libraries using error prone PCR. **b**, Retrotransposition assay and *Alu* integration into the genome with the help of L1 transposase machinery. The transposition vector contains RNAPol*III* 7SL driving an *Alu*-Neo cassette that is spliced and complexed with the ORF2 transposase machinery from the helper plasmid. The *Alu*-Neo cassette gets reverse transcribed and integrates randomly into the genome allowing the neomycin resistance gene to be expressed through the RNAPol*III* SV40 promoter thus conferring G-418 sulphate resistant colonies. **c,** Retrotransposed *Alu* resurrection and retrieval from the genome and sequence

library generation. *Alu*-Neo integrations were selected using neomycin specific and *Alu* specific primers that generate a 1kb PCR product if neomycin is spliced due to retrotransposition. *Alu* specific primers are then used to amplify the integrated *Alu*s which are then processed for sequencing.

Results

Retrotransposition assay optimization

We developed a massively parallel jumping assay (MPJA), based on a previous assay for individual *Alu* transposition[9]. In this assay, the retrotransposition cassette includes an *Alu* transcript that is driven by a 7SL Pol*III* enhancer (**Fig. 1b**). The 3' of the *Alu* transcript has a Pol*II* driven SV40-Neomycin (*Neo*) cassette cargo in antisense orientation before a Pol*III* termination signal, i.e. polyT tract. Upon Pol*III* *Alu* transcription, the resulting RNA gets spliced, due to the presence of an autocatalytic intron that is independent of Pol*II* splicing and engages with the L1-ORF2 transposon machinery, which is needed for its genomic integration. The integrated *Alu* copy loses 7SL and the Pol*III* polyA retains the Pol*II* SV40-spliced *Neo* cassette, providing neomycin resistance for colonies that undergo a retrotransposition event, which can then be used for selection.

We first set out to standardize the retrotransposition assay, using *Alu*Ya5, a highly active Alu element[9], as a positive control. A vector containing *Alu*Ya5 was co-transfected with L1-ORF2 into HeLa cells (**Fig. 1b**). These cells were re-plated once before starting G-418 selection 4-5 days post transfection. After three weeks of G-418 stable selection, sizable colonies were observed (**Fig. 1b**, **Extended Data Fig. 1** showing colony assay). Of note, we also observed a small number of colonies in the negative control (non-L1ORF2), likely due to random integration of the *Alu*Ya5 plasmid. These random unspliced product integrations could be easily detected by PCR primers unique to the vector and thus can be readily excluded from further processing. All colonies, both from *Alu*Ya5 and negative control, were further processed for genomic DNA extraction to examine *Alu* transposition. Upon splicing, the amplicon should be 500bp shorter than the unspliced 1.5kb band, allowing to distinguish random integration events versus transposition events (**Fig. 1c**). We found *Alu*Ya5 colonies to have a 500bp shorter PCR product compared to zero colonies from the negative control (**Extended Data Fig 2a**). This jumping assay is therefore extremely robust in detecting *Alu* sequences that lead to retrotransposition.

Selection of *Alu* elements for MPJA

Having optimized the retrotransposition assay, we next set out to choose a library of *Alu* elements to be tested by MPJA. A previous study that assessed the jumping activity of various *Alu* families from the human genome found the *Alu*J family to be inactive and have a highly divergent 280bp core region, the *Alu*S family to be moderately active, and the *Alu*Y family to be highly active, both having an intact 280bp core[4]. For our MPJA, we wanted to test what nucleotide changes are required to reactivate inactive *Alu* elements that have an intact 280bp core. We selected members of the *Alu*S family, since AluS family of retrotransposons largely consists of inactive elements with intact 280bp core and less diverged from the consensus AluSx sequence that is tested to be active in the retrotransposition assay[4]. A previous study that assessed 26 full length *Alu*S elements, found 21 to be inactive having low to zero activity[4]. Building on these results, we applied

stringent filters of inactivity to the same subfamily, to select three *Alu*S inactive candidates that have an intact 280bp core region and at least 95% sequence similarity with the consensus active *Alu*Sx sequence: *Alu*S-6b, *Alu*S-14b and *Alu*S-h1.1 (**Fig. 2a**). To further validate that they are indeed inactive or have very low jumping potential, we tested them individually in our assay confirming their activity as previously shown in a retrotransposition assay[4] (**Extended Data Fig 3**).

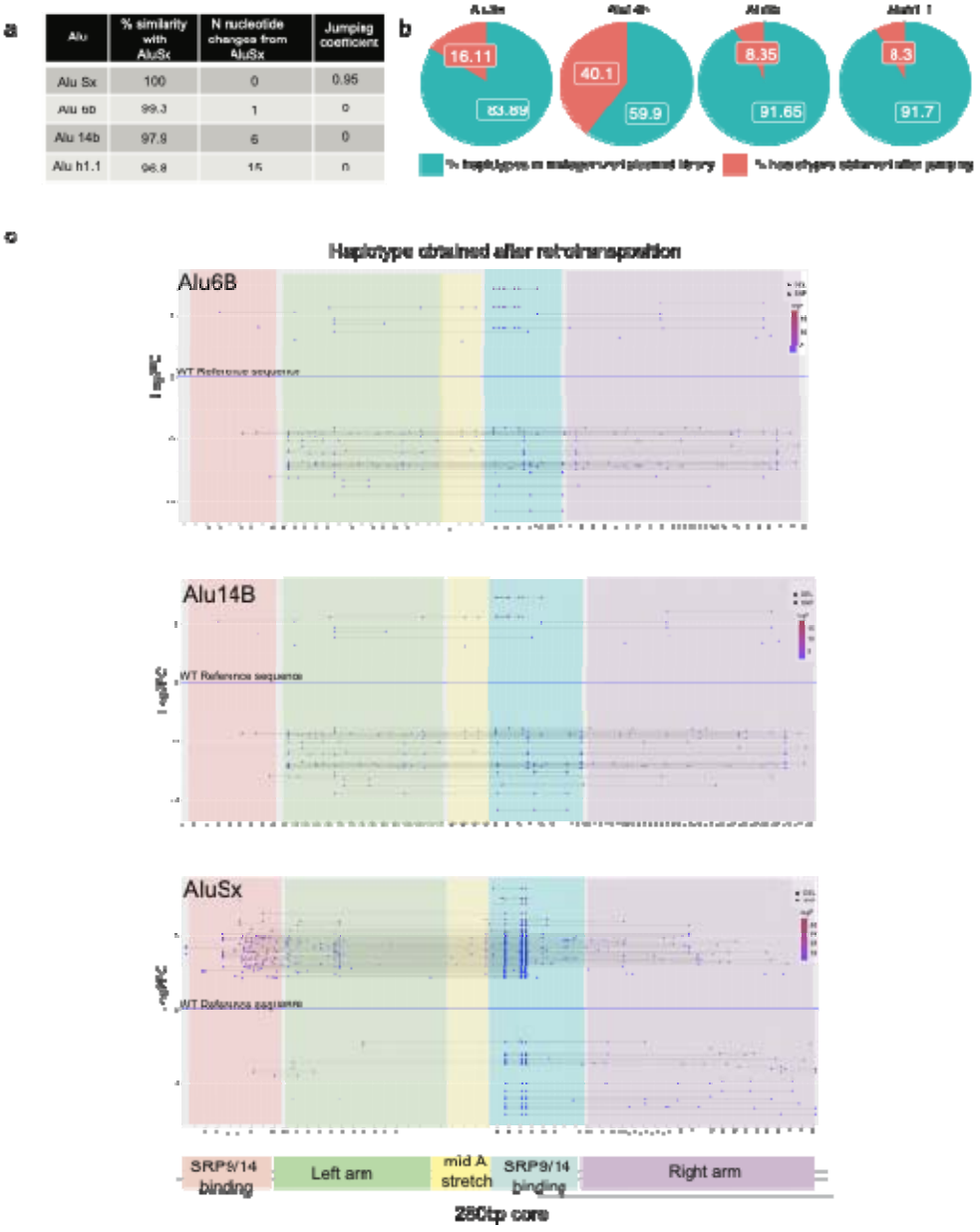

Figure 2: *Alu* selection and variants calling in the mutagenized library.
a, Sequence similarity (percent sequence identity), the number of sequence differences and the relative jumping potential to *Alu*Ya5 (positive control element considered to have activity score of 1) for all four tested *Alu* sequences. **b**, Pie charts showing the percent of haplotypes detected in the *Alu*-Mut jumping libraries from total haplotypes detected in *Alu*-Mut plasmid libraries. **c**, Variant calling at each position across the 280bp full-length Alu sequence in the *Alu*-mutagenized jumping/plasmid library and haplotype calling depicting only significant high jumping and low jumping haplotypes beyond a cut-off of +/- 2.5 log2 fold-change (Log2FC). Significant jumping was defined via the DESeq2 package using a Wald-test p-

value threshold of 10^{-5}. Lowest panel shows a schematic depicting the major structural features in *Alu*-RNA.

Alu-MPJA saturation mutagenesis

We synthesized all four *Alu* sequences and cloned them into the *Alu* jumping vector followed by Sanger sequence validation. We then used an error-prone PCR system that is estimated to generate a mutation every 1-16bp per kilobase (kb) of DNA (see Methods) to generate roughly 1-6 mutations per 300bp *Alu* element[10,11]. The mutagenized PCR product of each *Alu* element was then re-cloned into the *Alu* jumping assay vector to generate four mutagenized libraries: *Alu*Sx-mut, *Alu*6b-mut, *Alu*14b-mut and *Alu*h1.1-mut (**Supplementary Table 1**). To assess their complexity, libraries were subjected to two rounds of massively parallel sequencing, finding both replicates to show a high degree of overlap and correlation in count representation of identical variants (Pearson and Spearman correlation coefficiencies, *Alu*Sx p-rho 1.00 s-rho 0.8; *Alu*6b p-rho 1.00 s-rho 0.72; *Alu*14b p-rho 1.00 s-rho 0.77; *Alu*h1.1 p-rho 1.00 s-rho 0.77; **Extended Data Fig. 4 a**). We found all four libraries to be highly saturated for single nucleotide variants (SNVs) with frequencies of transitions, transversion and indels at each position (**Supplementary Table 2**, **Extended Data Fig. 5**). On average, we detected 10 variants in each individual *Alu* sequence, due to a small subset of variants that were fixed in the early error prone PCR amplification, leading to a 'founder effect' (**Extended Data Fig. 5**).

We next performed the retrotransposition assay with all four *Alu*-Mut libraries. Libraries were transfected into HeLa cells which were then treated with G-418 three days post transfection. Following 3-4 weeks of G-418 selection, colonies were amplified and replated to achieve confluency and later subjected to genomic DNA extraction. Using specific PCR primers (**Supplementary Table 3**), we amplified all the *Alu* sequences from the genomic DNA that were retrotransposed. First, we specifically amplified the 1kb spliced Neo-*Alu* amplicon to differentiate from any plasmid integrations that could happen without retrotransposition, allowing to resurrect only transposed *Alu* elements from the genome (**Extended Data Fig 2b)**. We also checked for primer set specificity of the Neo-Alu amplicon primers for any non-specific amplifications from the genomic DNA of un-transfected HeLa cells (**Extended Data Fig 2c)**. Next, we used primer targeting the vector sequence on 3' of the cloned *Alu* sequence and 5' *Alu* specific primers to amplify 300 bp fragments containing 280bp *Alu* core sequence (Methods). We then generated a sequencing library from the resulting ~280bp band and indexed and pooled libraries in equimolar ratios for sequencing.

Alu-MPJA saturation mutagenesis analysis

We next analyzed the *Alu*-Mut library (the plasmid variant library) and *Alu*-jumping library (the library following retrotransposition) for each *Alu* element. We first obtained the total number of read counts for single nucleotide variants in each plasmid (*Alu*-Mut) and *Alu*-jumping library. We observed that most of the variants fall within the mismatch window of up to <10 nucleotides changes for all four *Alu* elements (**Extended Data Fig. 6**) in both *Alu*-Mut and *Alu*-jumping libraries. We then calculated log2 fold-change

(log2FC) based on the frequency of reads from *Alu*-jumping vs *Alu*-Mut library. The significant variants with a positive fold-change and negative fold-change were determined using the DESeq2 R-package[12] (**Extended Data Fig. 5**). We normalized fold-changes with the measured activity of the reference *Alu* sequence that we started our mutagenesis on to provide a 'jumping score' for the sequences (**Fig 2a**). We observed that highly enriched SNVs in the *Alu*-Mut plasmid library, either due to PCR bias or founder effect, were not necessarily enriched in the *Alu*-Mut jumping library, confirming that our measurements are focused on jumping activity (**Fig. 2c**).

Next, we compared results from the two *Alu*-Mut jumping technical replicates (**Extended Data Fig. 4b**), finding them to show good correlations (Pearson and Spearman correlation coefficients, *Alu*Sx p-rho 1.00 s-rho 0.59; *Alu*6b p-rho 0.95 s-rho 0.60; *Alu*14b p-rho 1.00 s-rho 0.54) except for *Alu*h1.1, which had a poor correlation between replicates (*Alu*h1.1 p-rho 0.44 s-rho 0.16) (**Extended Data Fig. 4b**). We thus decided to remove *Alu*h1.1 from all subsequent analysis. We also screened our *Alu*-Mut plasmid libraries for the possible phenomenon of index hopping reported earlier on Illumina sequencing platforms[13]. Despite using dual-indexing, which was reported previously to be able to address the inaccuracies of multiplexing, we did observe a low proportion of potential index-hopping (for *Alu*Sx 0.8%, *Alu*6b 2.3%, *Alu*14b 3%, total 2.2% in *Alu*-Mut plasmid libraries)[14]. We note that this might also be caused by spurious sharing of sequences or could be the result of cross-contamination before library prep, and we removed all these reads in our subsequent analysis.

We annotated *Alu* haplotypes, requiring a limit of up to 10 nucleotide changes per element, as we wanted to focus on the minimum number of sequence variants required to change the activity of the *Alu* element. In addition, we wanted to prioritize those haplotypes in line with the low rate of sequence variants created by the mutagenesis strategy that we adopted, which typically allows 1-16 base changes per kb according to the Genemorph mutagenesis kit (see Methods). We called a haplotype only if it had read counts higher than two in the plasmid library and greater than ten in the jumping library. In total, we annotated 20,733, 48,641 and 35,849 haplotypes in *Alu*Sx, *Alu*6b and *Alu*14b respectively in the plasmid library. Of these 9,493, 10,463 and 32,176 haplotypes were found in the jumping libraries of *Alu*Sx, *Alu*6b and *Alu*14b, respectively (**Fig. 2**; **Supplementary Tables 4-7**). These numbers include both high and low jumpers without any further cutoffs. We found that the aforementioned highly enriched SNVs, which likely appeared either due to PCR bias or founder effect during library generation, were not significantly enriched in Alu jumping libraries (with +/-2 log2FC, p-value < 10^{-5}), suggesting that they have a minimal effect on jumping activity (**Fig. 2c**). Interestingly, we also observed in the in *Alu*-jumping libraries some fragments (>20-25 Mismatches) that were shorter than 200bp that were missing the right arm monomer (**Extended Data Fig. 7**), fitting with previous reports that showed that *Alu* elements missing this arm can still retain jumping activity[15]. We checked where the reads start (**Extended Data Fig. 8**) and all these <200bp fragments were clustered at around the same position in the *Alu* after the left monomer (**Extended Data Fig. 7**). We confirmed via PCR amplification using *Alu* recovery primers (**Supplementary Table 3**) that these short reads are not an artifact due to non-specific amplification from the untransfected

HeLa cell genome (**Extended Data Fig. 2c)**. We removed these haplotypes from our subsequent analyses.

Identification of haplotypes that alter *Alu* jumping potential

We next set out to identify *Alu* haplotypes that lead to a significant increase in retrotransposition, by measuring differential enrichment in *Alu*-Mut jumping vs *Alu*-Mut plasmid libraries (**Fig. 2c; Fig. 3a-c**). For each library, the normalized counts of all haplotypes were calculated using DESeq2[12]. We estimated log2FC by comparing Jumping haplotypes to Plasmid library (Methods) and normalizing to the reference Alu sequence (**Fig. 3a-c**). We applied a cuttoff of Log10P p-value threshold of 10^{-5} for significant haplotypes and refer to >2 log2FC as high jumpers and <2 log2FC as low jumpers respectively (**Fig. 3a-c**). Haplotypes that were observed in the plasmid library but not in the jumping library, we refer to as non-jumpers. We observed that majority of the unclassified haplotypes that did not fall under any of the above classes (**Fig. 3a-b**) for *Alu*6b and *Alu*14b have close to or lower log2FC than their reference sequence signifying their inherent low jumping potential. In contrast, for *Alu*Sx, which is inherently active, there were more non-classified haplotypes showed higher log2FC (**Fig 3c**). We next analyzed the number of nucleotide changes required to obtain jumping activity. We found that a minimum of two changes are needed for *Alu*6b and five for *Alu*14b. For *Alu*Sx, as it is already an active jumper, we analyzed how many changes are needed to increase its jumping potential, finding that one nucleotide change was sufficient to substantially increase the >2 log2FC (**Fig. 3d**).

Our assay only detects variants that are present in *Alu* sequences that jumped but does not detect haplotypes that are either non-jumpers or could be absent due to other reasons that are not related to retrotransposition ability. We thus applied a significantly higher threshold of read counts for potential haplotypes that reduce jumping activity, requiring over 50 reads in the plasmid and zero reads in the jumping library in any of the replicates. This allowed us to separate out potential non-jumpers (**Fig. 3a-c**) from background that could be due to various experimental effects. Combining results of the low and non-jumpers, we found that one nucleotide change for *Alu*Sx, two nucleotide changes for *Alu*6B and seven nucleotide changes for *Alu*14B were required to decrease or abolish jumping potential (**Fig. 3d**).

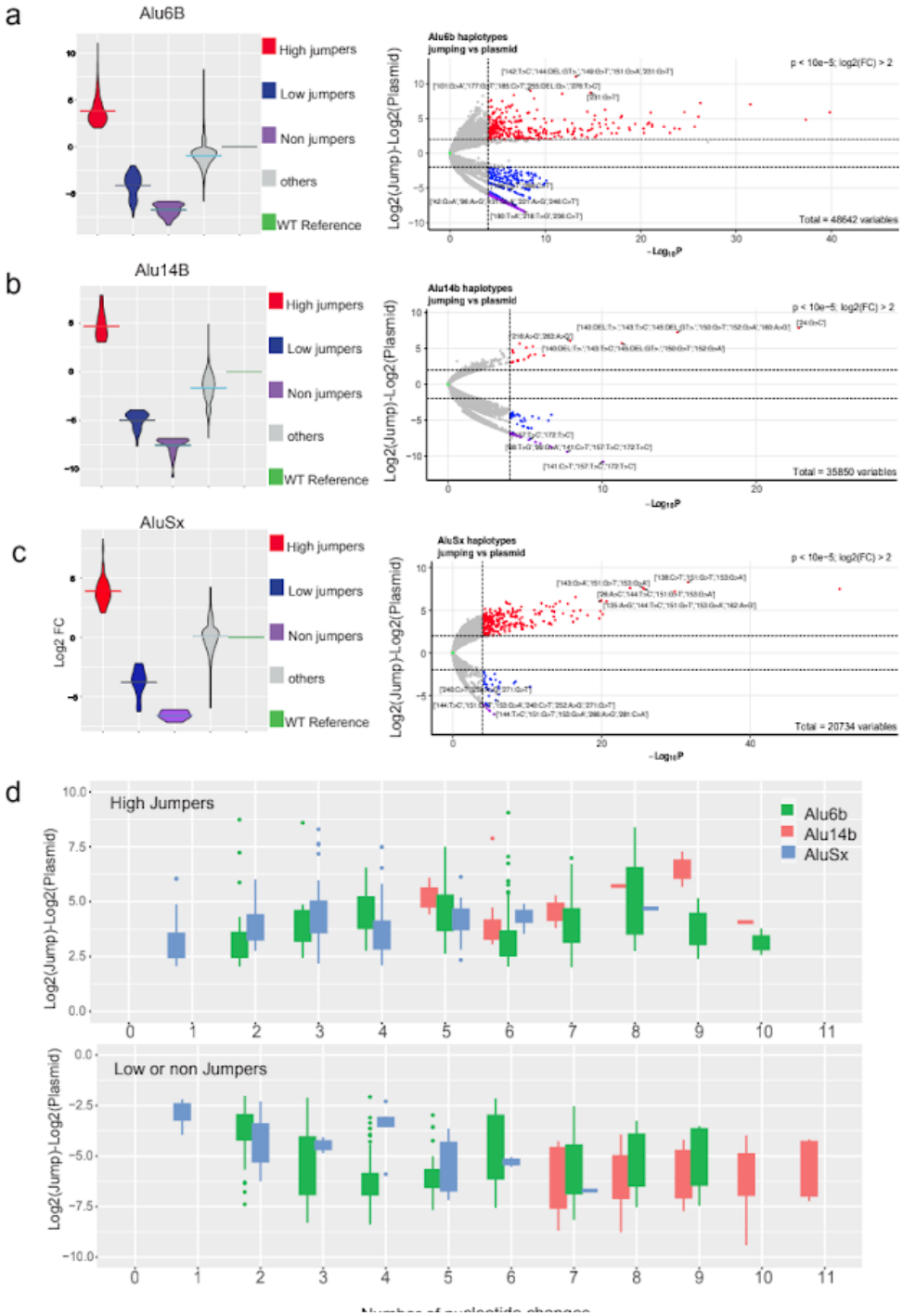

Figure 3: Haplotypes of dominant jumping effects.
a-c, Violin (left panel) and Volcano (right panel) plots showing fold change differences in high jumpers, low jumpers and possible non-jumping *Alu* haplotypes with cutoff of +/- 2 log2 fold change (Log2FC) for *Alu*14B-Mut (**a**), *Alu*6B-Mut (**b**), *Alu*Sx-Mut (**c**), respectively. Significant effects were defined using the DESeq2 package with a Wald-test and p-value threshold of 10^{-5}. Classes were defined as significant high jumpers in red (Log2FC > 2), significant low jumpers in blue (Log2FC < -2, jumping counts > 0) and non-jumpers in purple (Log2FC < -2, jumping counts = 0 and plasmid count > 50). Non-significant haplotypes are shown in grey. Plots are normalized to a zero Log2FC of the reference or wildtype sequence (in green) of each element. **d**, The number of nucleotide changes observed in the library with respect to the reference or wildtype sequence for high jumper (top) and low/non jumper haplotypes (bottom).

Mutations that affect jumping are associated with SRP binding domains

We next analyzed the dataset along the length of the *Alu* element to reveal positions or domains where mutations have potential to contribute towards positive or negative jumping activity. We used a 5bp sliding window analysis to score the activity from different nucleotide variants over the sliding window across the full-length of the *Alu* element (**Fig. 4**). We fit multiple linear regression models and plot the coefficients reflecting the combined effect of variants over the respective 5bp tiles/windows. Our results indicate that the log2FC haplotype activity presented above correlates well with the 5bp window analysis, which allows us to compensate for the differences in variant density and representation of certain variants from the saturation mutagenesis process.

Alu retrotransposons are transcribed by the RNA Pol*III* machinery and fold into a specific RNA structure[16]. This secondary structure of *Alu*-RNA is crucial to engage the transposition machinery[15,17]. *Alu*-RNA has distinct regions in its left half (1-120bp), termed the left arm or monomer, followed by a middle A-stretch and a right arm or monomer (150-300bp) (**Fig. 2c**). The left and right arms of *Alu*-RNA fold independently in a series of stems and loops. One of the secondary structures that is formed at the 5' of each left and right arm is the bipartite stem-loop structure that binds to the signal recognition particle (SRP9/14) complex. The SRP binding site on left arm has annotated BoxA and BoxB regions that are known to be crucial for retrotransposition[6,18,19]. SRP complex binding is crucial for the process of transposition and nucleotide changes that alter the binding affinity of the SRP complex can affect the transposition efficiency[17]. The 5bp window analysis clearly indicated an enrichment of activity alteration in the SRP binding domains that correspond to *Alu* RNA secondary structure.

We used RNAfold to predict the secondary structure of the *Alu*-RNA[20]. We analyzed whether nucleotide changes in regions that were predicted to positively affect the haplotype frequency are enriched in the SRP binding stem-loop structure (**Fig. 4**). We performed a hypergeometric test of variants in high jumpers and low/non jumpers in the SRP binding regions for both left and right SRP binding structures. Interestingly, we observed that variants in the SRP binding stem-loop structures were highly enriched for high jumpers compared to low/non jumpers in all *Alu* classes tested (Hypergeometric test *Alu*6b P value=0.00, Alu14b P value=1.65e-05, *Alu*Sx P value=0.00) (**Fig. 4**). We performed sequence analysis on both left and right SRP binding structures to pinpoint changes that are contributing towards Alu jumping potential (**Fig. 4**, **Extended Data Fig 9**). We found that individual nucleotide alterations in these regions were sufficient to modify *Alu* jumping activity. Furthermore, we observed clear differences in the frequency of certain nucleotides at specific positions in high jumping vs low jumping haplotypes. In summary, our results show that nucleotide changes overlapping predicted *Alu* SRP binding regions have a major effect on transposition.

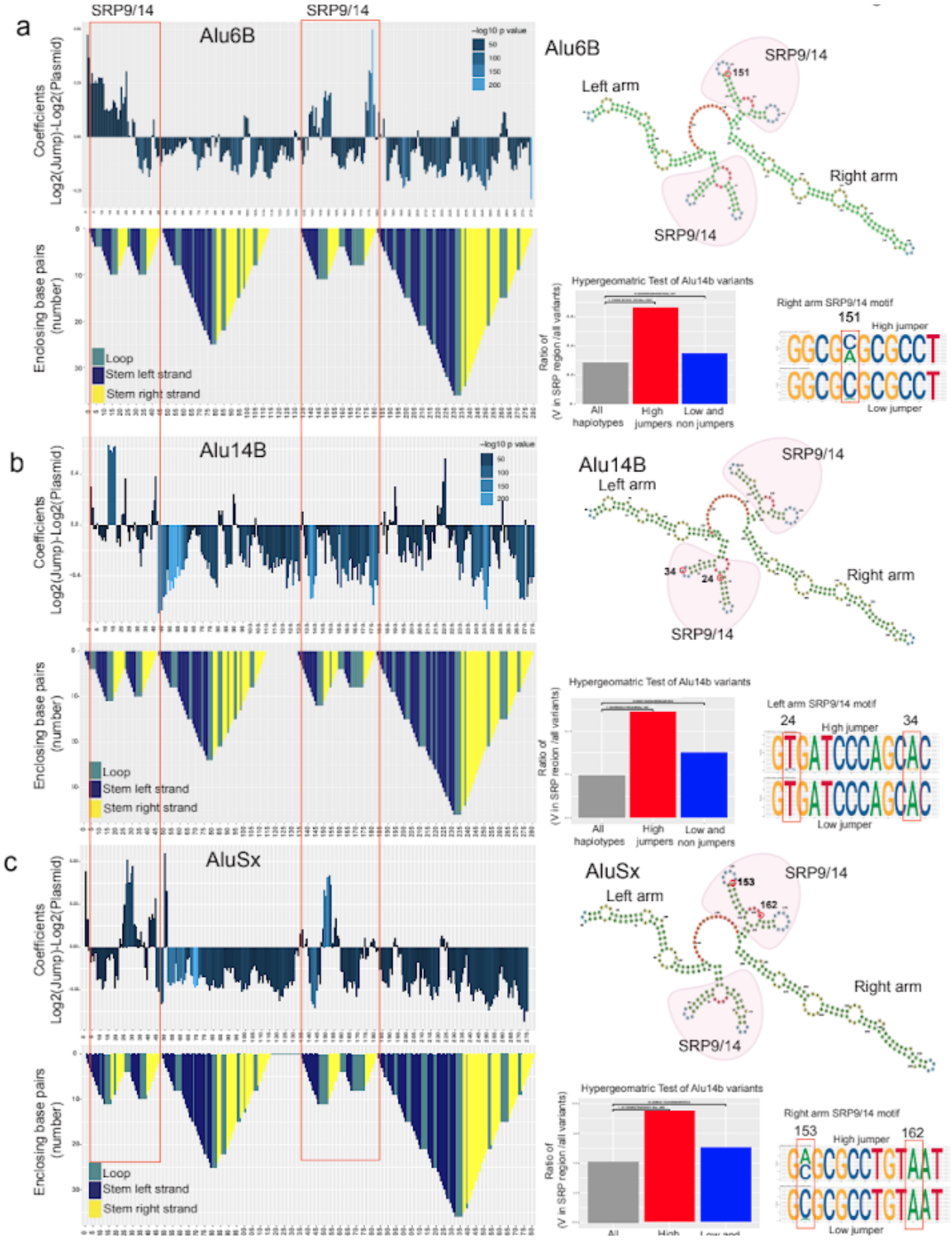

Figure 4: Mutations that affect jumping are associated with SRP binding domains.
a-c, In the top panel 5bp-sliding window analyses of *Alu*6B (**a**), *Alu*14b (**b**), and *Alu*Sx (**c**), are shown along with the *Alu*-RNA folding structure in the panels below. The folding structure is reverse mountain coded with different colors for each secondary structure: stem loop-unhybridized in teal, hybridized left strand of the stem in dark blue and hybridized right arm of the stem in yellow. The upper right panel show predicted RNA secondary structures with the left SRP and right arm SRP binding folds highlighted in red. The right lower panels show hypergeometric analyses with the enrichment of variants of dominant jumping effect in SRP binding regions (red bar) versus the enrichment of variants of negative jumping effect in SRP binding regions (blue bar). Next to these, sequence logos show DNA nucleotide composition observed in high jumper (top) and low jumper (bottom) haplotypes at select positions.

Comparison to *Alu*S sequences in the human reference genome

We next set out to quantify how many mutations can lead to the activation of *Alu*S elements in the human reference genome. The human genome has around half a million (551,383) copies of *Alu*S elements and around 852 of those have an intact 280bp core[4]. We extracted *Alu*S elements previously identified from the genome[4] and added a custom identifier for each (**Supplementary Tables 8-9**). We first aligned our *Alu*S jumping haplotypes to the subset of *Alu*S sequences in the human genome that have an intact 280bp core region, finding that none of these haplotypes have an exact match with these elements. By testing 27 *Alu*S elements individually, it was previously shown that no full length *Alu*S is active in a retrotransposition assay when it is 10% (28 nucleotide changes) divergent from the consensus active AluSx sequence[4]. We first plotted the total number of haplotype counts generated in the Alu-Mut jumping libraries by the number of mismatches to existing AluS sequences in the human genome with an intact 280bp core (**Fig. 5a**). From our MPJA study, we found that the closest jumping haplotypes for Alu6B or AluSx are seven mismatches away from endogenous AluS and ten mismatches for Alu14B (**Fig. 5a-b**). Analyzing only the high jumper haplotypes, we found eight mismatch differences for Alu6B or AluSx and twelve mismatches for Alu14B (**Fig. 5b**). We also looked at how many AluS elements in the human genome could potentially become jumpers upon mismatches. We plotted the number of genomic AluS that have mismatches with only high jumper haplotypes from our mutagenesis library. We found up to fifteen high jumping haplotypes in our library similar to eight distinct Alu elements from the genome that are 8-16 mismatches away from being high jumpers (**Fig. 5c**).

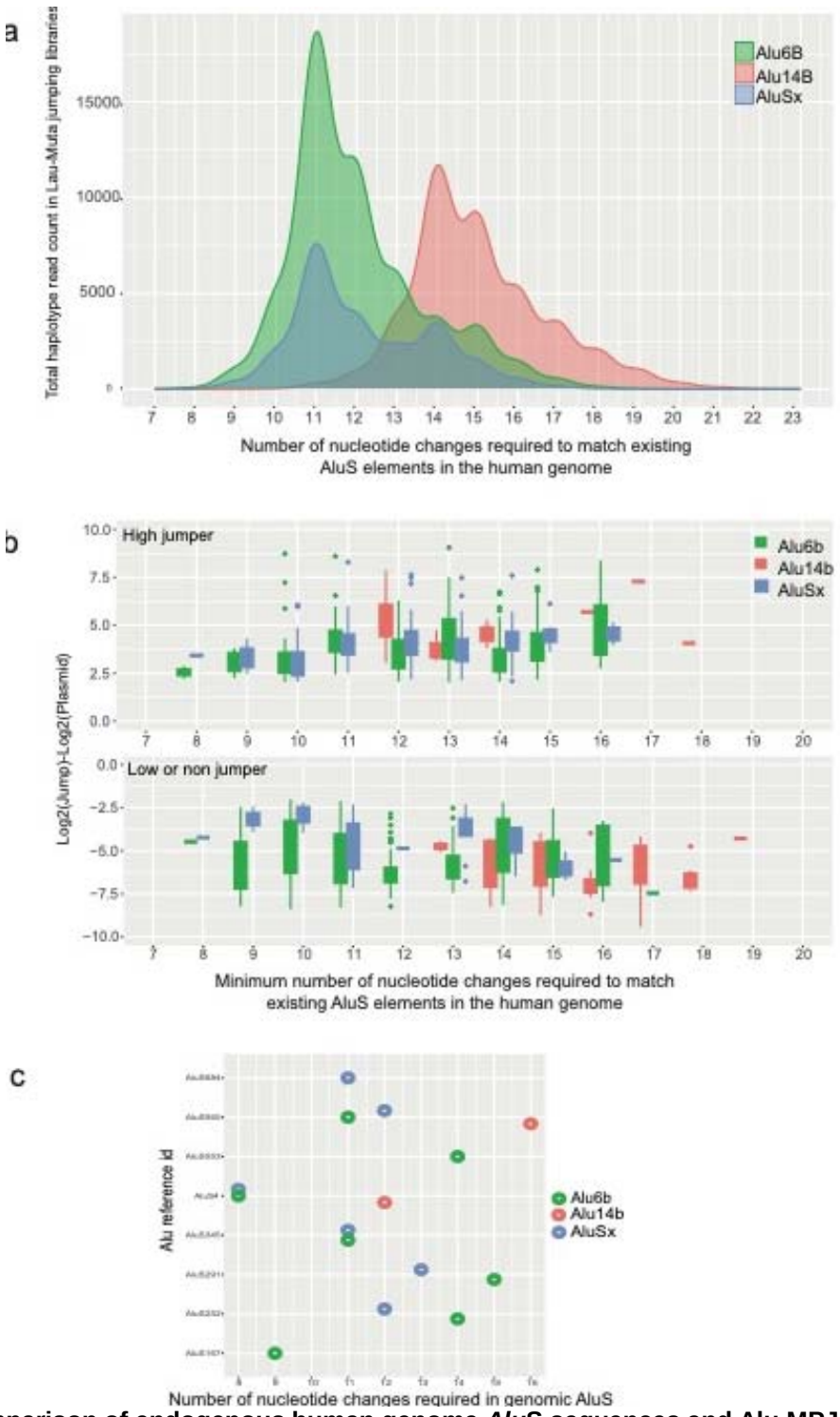

Figure 5: Comparison of endogenous human genome *Alu*S sequences and Alu MPJA haplotypes. **a**, 852 genomic *Alu*S from the human genome were compared to the haplotypes found in the mutagenized libraries of each element. We detected haplotypes that are at least seven mismatches away from the genomic *Alu*S in *Alu*Sx and *Alu*6B mutagenized libraries and at least eleven mismatches away for Alu14B haplotypes. **b**, Bar chart showing the mismatch distance of high (top panel) or low (bottom panel) jumpers in the MPJA from genomic AluS sequences. **c**, Dot plot showing fifteen identified *Alu*S elements from the human genome (**Supplementary Table 8**) that require minimum number of nucleotide changes to match with the high jumper haplotypes.

Discussion

Retrotransposition of genomic elements is a major force driving genomic diversity[1]. Around 11% of the human genome consists of *Alu* retrotransposons with 1.4 million copies existing in various genomic locations. With a germline mutation rate of 1.2×10^{-8} per base pair per generation[21–23], the sequences of many *Alu*s can change over time, which could potentially lead to their reactivation. Most *Alu* elements are inactive due to sequence disintegration, but a few have maintained sequence features necessary for jumping, i.e., an intact 280bp core sequence with BoxA and BoxB in the left monomer as well as a PolyA tail[6]. It is unknown how nucleotide changes present throughout the *Alu* elements may affect their jumping potential. The assays developed so far measure the activity of an individual transposon precluding our ability to identify nucleotide changes at a higher resolution level that could affect their jumping potential. To address this, we developed an assay that can test in a high-throughput manner the jumping capability of thousands of sequences. We used four different *Alu*S elements, constructing libraries containing 167,534 unique haplotypes and tested their jumping ability. This assay could be easily modified for other *Alu* families, for example the *Alu*Y family that is still active in the human genome, as well as other types of retrotransposons.

We found that it takes a minimum of one for *Alu*Sx, two for *Alu*6B and five for *Alu*14B nucleotide changes to increase the jumping potential, most of them residing in SRP binding domains. Here, we limited our analysis to ten mismatches, thus limiting our ability to observe effects for more sequence alterations. To overcome limitations of the error-prone PCR step, a synthetic design of the *Alu* library (e.g., by oligosynthesis) could incorporate widespread changes including larger deletions or insertions. In addition, designed Alu sequences could potentially also test for variant combination effects, i.e. analyzing numerous variant combinations to identify how they interact and affect each other.

There are several critical steps that can affect *Alu* retrotransposition. The *Alu*-RNA with 3'polyA tail folds and forms the secondary structure crucial for binding of SRP9/14 complexes. There is one SRP9/14 binding site in each arm of the Alu RNA. This ribonucleoprotein complex recruits L1-ORF2 machinery that has reverse-transcriptase and transposon complex and randomly integrates into the genome. Any disruption during this process can lead to abrogation of the jumping potential[9]. In our dataset, non-jumpers could have mutations that disrupt any of the following processes: 1) Transcription of the Alu element via RNA Polymerase III; 2) polyA tailing; 3) misfolding of Alu-RNA; 4) poor binding of the SRP9/14 complex; 5) inability to recruit the L1-ORF2 machinery. As this assay is primarily designed to identify jumping activity, it is limited in identifying variants that inhibit jumping. Despite that, using stringent conditions, we did detect a small number of variants that abolished jumping, although with lower confidence. Further studies will be required to parse out these other retrotransposition associated factors in a systematic manner.

We observed a spike in fragments that have >20-25 mismatches only in jumping libraries and not in plasmid libraries. We think this could be due to the inherent nature of our assay system/cells. Since we do not know how *Alu* RNA is processed in these cells, we filtered these fragments from our analyses. However, alternative cells or assay systems may not show this high mismatch number and/or provide an understanding of its cause. Surprisingly, we also observed a small proportion of fragments to be smaller than 200 nucleotides in length in both our jumping libraries and *Alu*-mut plasmid libraries, that could not be explained based on how we processed these sequencing libraries. Of note, it has been reported earlier that the minimal *Alu* RNA length required for jumping could be less than 200 bp and devoid of its right arm monomer[15], fitting with our finding. Here, we removed these sequences from our analysis, but future MPJA could be used to test what are the critical *Alu* sequences in this left arm monomer that allow it to jump.

MPJA could also be applied to investigate whether *Alu* activity could potentially be involved in evolutionary adaptations or human disease by testing specific somatic mutation haplotypes detected in patients. The haplotype data that we generated could be compared with patient data such as whole genome sequences from cancer patients, to elucidate if any observed mutations in *Alu* sequences could lead to its reactivation or retrotransposition and consequently increase genomic mutation frequencies. It is also important to consider that *Alu* elements need an active L1 to provide the retrotransposition machinery, as they are non-autonomous jumpers. L1 is active during germline development, early embryonic development, cancer progression or in neurodegenerative diseases which could affect *Alu* retrotransposition if inherently active *Alu* elements are present in the genome[24–31]. A previous study that individually examined the retrotransposition ability of several *Alu*S, estimated that at least 28 nucleotide changes (10% variation) from the consensus active sequence *Alu*Sx are needed to obtain jumping activity[4]. Here, we found that this number can be much lower, with up to seven mismatches for some of the *Alu*S subclasses. We focused on the *Alu*S family for our study, but future studies could be easily extended to other classes of Alu elements in the human genome, for example the youngest *Alu*Y family elements that are currently active in the genome and have even higher potential to be involved in human disease.

In summary, we provide a novel high-throughput approach to test the jumping activity of thousands of *Alu* retrotransposons in parallel that could be adopted for other retrotransposons including DNA transposons. This could facilitate the studies of jumping elements *en masse* in different conditions and be used to address various biological questions related to transposons.

Materials and methods

Retrotransposition assay

The Alu retrotransposition assay was performed as previously described[4,5,9]. Briefly, HeLa cells (a kind gift from Dr. Astrid Engel, Tulane University) were transfected with lipofectamine LTX reagent using pYa5-neo and pL1-ORF2 or L1 helper (kind gift from Dr. Astrid Engel, Tulane University) at a 1:1 concentration ratio. After 72 hours, DMEM with G-418 (500ug/ml) and pen-strep (100ug/ml) was replaced and the neomycin selection medium was refreshed every 72 hrs for 2-3 weeks. Colonies were washed with PBS and stained with 0.5% Crystal Violet staining solution containing 10% ethanol and 50% methanol.

Error prone PCR and Alu library cloning

*Alu*Sx, *Alu*14B, *Alu*6B and *Alu*h1.1 were synthesized as gene blocks (IDT) and cloned into BamHI-AatII sites in the pYa5-neo vector using Infusion cloning kit (Takara) following the manufacturer's protocol. To generate mutagenized libraries with error-prone PCR, we used the GeneMorphII random mutagenesis kit (Agilent) following the manufacturer's protocol. Two independent reactions with 10ng of plasmid subjected to 25 PCR cycles were pooled and cleaned using the PCR clean up kit (Qiagen; 28104). The mutagenesis primer (**Supplementary Table 1**) towards the 3' of the *Alu* matches the last base in the vector and therefore the 3' of *Alu* was subjected to random mutagenesis. To facilitate subsequent resurrection and retrieval of the jumping *Alu* haplotypes from the genome, mutagenesis was initiated only after the first 27 bases into the 5' end of the *Alu*. This allowed us to specifically resurrect the retrotransposed and integrated Alu-Neo cassette from the genome via nested PCR amplification strategy (**Fig. 1c**). Mutagenesis primers had 15bp overhangs at either side to aid Infusion cloning into PstI-AatII sites of the pYa5-neo vector. Four independent Infusion-cloning reactions were set up and transformed into four different vials of 150ul each of Stellar competent bacterial cells. Each recovered competent cell vial was used as inoculum for 75ml of LB broth which was cultured overnight. Four different plasmid midi preps (Midi prep kit; Qiagen) were pooled to obtain the *Alu*-Mut library. This procedure was performed separately for each *Alu* element.

Jumping Alu-Mut library retrieval

MPJA was performed on 15cm cell culture dishes (Corning). After stable selection for 3-4 weeks on G418, colonies were scrapped from plate for genomic DNA isolation using genomic DNA isolation kit (Promega). The spliced and retrotransposed Alu-Neo cassette (1kb) was isolated from genomic DNA using Alu recovery and resurrection primers described in **Supplementary Table 3.** This primer set (PCR-a) allowed to differentiate between the randomly integrated unspliced plasmid DNA product which is 1.5kb in length. The 1kb PCR fragment was then used for nested PCR-b using primers described in **Supplementary Table 3**. Plasmid mutagenesis libraries were prepared by using PCR-b on the mutagenized plasmid library as the template directly. The amplicon library was then prepped as described above using the Ovation low complexity kit. Libraries were sequenced on Illumina MiSeq with Paired End (PE) 250bp reads and multiple runs.

Library sequencing and analysis

Alu-Mut plasmid libraries were generated using PCR-based amplification with primers sets described in **Supplementary Table 3** using the Ovation low complexity library prep kit (Ovation® Library system for Low complexity samples Part no: 9092-16). Briefly, the PCR amplicon was end-repaired, and forward and reverse diversity adaptors were ligated, filled-in and dual index barcode i5 and i7 primers were used to perform PCR for 10 cycles to generate the Alu library fragments (**Supplementary Table 10**). Libraries were sequenced on Illumina MiSeq with multiple PE 250bp runs. PE reads were processed and stitched together to full-length fragments with PEAR[32] and aligned with BWA[33] to Alu reference sequences. Custom scripts were used in the analysis for variant calling and filtering. We identified unique haplotypes based on the alignment information stored in BAM format, specifically we used position and CIGAR fields as well as MD (differences to the aligned reference sequence) and NM (alignment edit distance) tags to identify identical haplotypes and to count their abundance. The edit distance in the NM field was used to filter and to report the minimum number of required edits to the reference sequence. We performed differential activity analysis using DESeq2[12] on unique haplotypes.

High, low, non-jumper definitions

For all haplotype variants, we compared the fragment counts from two jumping replicates and two plasmid replicates. We use DESeq2[12] to define high jumpers (log2 fold-change > 2, p-value < 10^{-5}), low jumpers (log2 fold-change < -2, p-value < 10^{-5}, jumping counts > 0) and non-jumpers (log2 Fold change < -2, p-value < 10^{-5}, jumping counts = 0, plasmid counts > 50).

Inferring variant level activity effects

To infer the activity effects of individual variants along the different Alu elements, we considered every variant (including deletion and insertion) along all haplotypes of a specific experiment. These were included in a matrix with the haplotypes as rows including their plasmid count, jumping count, and *N* binary columns indicating whether specific variants were associated with that haplotype. We then fit a combined multiple linear regression model of both replicates[10] (Equation 1).

$$log_2\left(Jumping_{i,j}\right)\left\{\begin{matrix} log_2\left(Plasmid_{rep1,j}\right) Vi=1 \\ 0 Velse \end{matrix}\right\}+\left\{\begin{matrix} log_2\left(Plasmid_{rep2,j}\right) Vi=2 \\ 0 Velse \end{matrix}\right\}+N+offset \qquad (1)$$

Inferring sliding-window activity effects

To infer the regional activity effects along the different *Alu* elements, we applied a 5bp sliding-window approach where we use 5bp windows with an offset of 1 bp along the length of the *Alu* elements. We considered every variant (including deletion and insertion) in each window and considered the combined plasmid count and jumping counts for the center of the window. We then fit a combined multiple linear regression model of both replicates[10].

Sequence analysis and logos

We analyzed the frequency of nucleotides in short sequence regions, i.e. motifs, by contrasting haplotypes of high and low jumping activity. For both SRP binding regions (left arm and right arm in each element), we generated logo plots (ggseqlogo R package[34]) with input of SRP left and right binding regions respectively.

Hypergeometric test for SRP binding enrichment

To determine whether variants in SRP binding regions are overrepresented or underrepresented in high, low or non-jumpers, we took the total variants from all *Alu* elements (including all haplotypes) as background. Similar to the gene set enrichment analysis, a significant p-value obtained from the hypergeometric test suggested that the variants in SRP binding regions are enriched in high jumpers, whereas a non-significant p-value suggested that variants in SRP binding regions are not enriched in low or non-jumpers. The p-value was adjusted for multiple testing to control the false discovery rate (Benjamini Hochberg correction).

RNA structure prediction

We used the RNAfold webserver[20] to predict the secondary structures of single stranded RNA sequences, for which we first converted *Alu* DNA sequences to RNA. We manually set folding constraints to obtain an *Alu*-like structure, i.e. we are not just reporting the minimum free energy structures, but an *Alu* RNA like structure that requires secondary structure folding of left arm monomer and right arm monomer separately with unstructured middle A stretch.

Data availability

All sequencing data for this study has been deposited in the GEO database under accession code PRJNA1098913.

Acknowledgments: This work was supported in part by the National Human Genome Research Institute (NHGRI) grant numbers UM1HG009408 (NA) and UM1HG011966 (MK, NA) and National Institute of General Medical Sciences R01GM142112 (NA), Innovative Genomics Institute, IGI-RIDER and IGI-WIES funding (NM). We thank Astrid Engel for providing valuable guidance for retrotransposition assays and sharing vectors and HeLa cell line.

Authors contributions: NM, NA conceptualized the study and wrote the manuscript. MK, JZ conceived the computational analysis. JZ, MK, NM performed analyses and created figures. JM, LC provided guidance for selection of Alu candidates. NM, AS, LD, ZL, SR, JS, YH performed the experiments.

References

1. Cordaux, R. & Batzer, M. A. The impact of retrotransposons on human genome evolution. *Nature Reviews Genetics* vol. 10 Preprint at https://doi.org/10.1038/nrg2640 (2009).
2. Cordaux, R., Hedges, D. J., Herke, S. W. & Batzer, M. A. Estimating the retrotransposition rate of human Alu elements. *Gene* **373**, (2006).
3. Deininger, P. Alu elements: Know the SINEs. *Genome Biology* vol. 12 Preprint at https://doi.org/10.1186/gb-2011-12-12-236 (2011).
4. Bennett, E. A. *et al.* Active Alu retrotransposons in the human genome. *Genome Res* **18**, (2008).
5. Dewannieux, M. & Heidmann, T. Role of poly(A) tail length in Alu retrotransposition. *Genomics* **86**, (2005).
6. Roy-Engel, A. M. *et al.* Non-traditional Alu evolution and primate genomic diversity. *J Mol Biol* **316**, (2002).
7. Kopera, H. C. *et al.* LEAP: L1 element amplification protocol. in *Methods in Molecular Biology* vol. 1400 (2016).
8. Streva, V. A. *et al.* Sequencing, identification and mapping of primed L1 elements (SIMPLE) reveals significant variation in full length L1 elements between individuals. *BMC Genomics* **16**, (2015).
9. Dewannieux, M., Esnault, C. & Heidmann, T. LINE-mediated retrotransposition of marked Alu sequences. *Nat Genet* **35**, (2003).
10. Kircher, M. *et al.* Saturation mutagenesis of twenty disease-associated regulatory elements at single base-pair resolution. *Nat Commun* **10**, (2019).
11. Lee, S. O. & Fried, S. D. An error prone PCR method for small amplicons. *Anal Biochem* **628**, (2021).
12. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* **15**, (2014).
13. Illumina. Effects of Index Misassignment on Multiplexing and Downstream Analysis. *Illumina* (2017) doi:10.1101/125724.
14. Kircher, M., Sawyer, S. & Meyer, M. Double indexing overcomes inaccuracies in multiplex sequencing on the Illumina platform. *Nucleic Acids Res* **40**, (2012).
15. Ahl, V., Keller, H., Schmidt, S. & Weichenrieder, O. Retrotransposition and Crystal Structure of an Alu RNP in the Ribosome-Stalling Conformation. *Mol Cell* **60**, (2015).
16. Huck, L. *et al.* Conserved tertiary base pairing ensures proper RNA folding and efficient assembly of the signal recognition particle Alu domain. *Nucleic Acids Res* **32**, (2004).
17. Weichenrieder, O., Wild, K., Strub, K. & Cusack, S. Structure and assembly of the Alu domain of the mammalian signal recognition particle. *Nature* **408**, (2000).
18. Häsler, J. & Strub, K. Alu elements as regulators of gene expression. *Nucleic Acids Res* **34**, (2006).
19. Conti, A. *et al.* Identification of RNA polymerase III-transcribed Alu loci by computational screening of RNA-Seq data. *Nucleic Acids Res* **43**, (2015).
20. Gruber, A. R., Lorenz, R., Bernhart, S. H., Neuböck, R. & Hofacker, I. L. The Vienna RNA websuite. *Nucleic Acids Res* **36**, (2008).

21. Campbell, C. D. & Eichler, E. E. Properties and rates of germline mutations in humans. *Trends in Genetics* vol. 29 Preprint at https://doi.org/10.1016/j.tig.2013.04.005 (2013).
22. Kong, A. *et al.* Rate of de novo mutations, father's age, and disease risk. *Nature* **488**, (2012).
23. Conrad, D. F. *et al.* Variation in genome-wide mutation rates within and between human families. in *Nature Genetics* vol. 43 (2011).
24. DiRusso, J. A. & Clark, A. T. Transposable elements in early human embryo development and embryo models. *Current Opinion in Genetics and Development* vol. 81 Preprint at https://doi.org/10.1016/j.gde.2023.102086 (2023).
25. González-Rico, F. J. *et al.* Alu retrotransposons modulate Nanog expression through dynamic changes in regional chromatin conformation via aryl hydrocarbon receptor. *Epigenetics Chromatin* **13**, (2020).
26. Larsen, P. A. *et al.* The Alu neurodegeneration hypothesis: A primate-specific mechanism for neuronal transcription noise, mitochondrial dysfunction, and manifestation of neurodegenerative disease. *Alzheimer's and Dementia* vol. 13 Preprint at https://doi.org/10.1016/j.jalz.2017.01.017 (2017).
27. Polesskaya, O. *et al.* The role of Alu-derived RNAs in Alzheimer's and other neurodegenerative conditions. *Med Hypotheses* **115**, (2018).
28. Larsen, P. A., Hunnicutt, K. E., Larsen, R. J., Yoder, A. D. & Saunders, A. M. Warning SINEs: Alu elements, evolution of the human brain, and the spectrum of neurological disease. *Chromosome Research* **26**, (2018).
29. Di Ruocco, F. *et al.* Alu RNA accumulation induces epithelial-to-mesenchymal transition by modulating miR-566 and is associated with cancer progression. *Oncogene* **37**, (2018).
30. Tang, R. B. *et al.* Increased level of polymerase III transcribed Alu RNA in hepatocellular carcinoma tissue. *Mol Carcinog* **42**, (2005).
31. Hwang, T. *et al.* Genome-wide perturbations of Alu expression and Alu-associated post-transcriptional regulations distinguish oligodendroglioma from other gliomas. *Commun Biol* **5**, (2022).
32. Zhang, J., Kobert, K., Flouri, T. & Stamatakis, A. PEAR: A fast and accurate Illumina Paired-End reAd mergeR. *Bioinformatics* **30**, (2014).
33. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, (2009).
34. Wagih, O. Ggseqlogo: A versatile R package for drawing sequence logos. *Bioinformatics* **33**, (2017).